S
363.728
H2rprc
1995
Bishop, Richard C
Peer review of
"Contingent
valuation of
natural resource
damages due to
injuries to the

ATE OF MONTANA
SOURCE DAMAGE PROGRAM

REVISED PEER REVIEW OF "CONTINGENT VALUATION OF NATURAL RESOURCE DAMAGES DUE TO INJURIES TO THE UPPER CLARK FORK RIVER BASIN" AND REBUTTAL REPORT

PREPARED BY:
RICHARD C. BISHOP

OCTOBER 1995

STATE DOCUMENTS COLLECTION
DEC 20 1995
MONTANA STATE LIBRARY
1515 E. 6th AVE.
HELENA, MONTANA 59620

S
363.728
H2rprc
1995

Bishop, Richard C
Peer review of "Contingent valuation of natural resource damages due to injuries to the

DATE	ISSUED TO
JUN 03 2997	

DEMCO 32-209

MONTANA STATE LIBRARY
363.728 H2rprc 1995 c.1 Bishop
eer review of "Contingent valuation of
3 0864 00096627 8

(9)

Revised October 1995

PEER REVIEW OF "CONTINGENT VALUATION OF NATURAL RESOURCE DAMAGES DUE TO INJURIES TO THE UPPER CLARK FORK RIVER BASIN"

STATE OF MONTANA
NATURAL RESOURCE DAMAGE LITIGATION PROGRAM

Prepared by:

Professor Richard C. Bishop
Department of Agricultural Economics
University of Wisconsin
Madison, WI 53706

October 25, 1995

INTRODUCTION 1

OVERALL VALIDITY OF THE CONTINGENT VALUATION METHOD 6

CONTENT VALIDITY ASSESSMENT 18

CONSTRUCT VALIDITY ASSESSMENT 42

CRITERION VALIDITY ASSESSMENT 50

COMPARISON OF CLARK FORK STUDY WITH TWO OTHER STUDIES 53

COMMENTS ON THE SCIENTIFIC STATUS OF CONTINGENT VALUATION 66

SUMMARY AND CONCLUSIONS 69

REFERENCES 73

LIST OF TABLES

Table 1: Content Validity Scores for the Clark Fork Study 30

Table 2: Comparative Content Validity Ratings for the Wilderness Area (WA), Clark Fork (CF), and Exxon Valdez Oil Spill (EV) Studies 53

Appendix A: Measuring Nonuse Values: Theory and Empirical Applications

Appendix B: Assessing the Content Validity of Contingent Valuation Studies

Addendum: Responses to Criticisms of My Peer Review by ARCO's Experts

INTRODUCTION

In a January 1995 report entitled "Contingent Valuation of Natural Resource Damages Due to Injuries to the Upper Clark Fork River Basin," a team from RCG/Hagler Bailly (now Hagler Bailly Consulting, Inc.) reported the results of a study they had conducted to value, in monetary terms, complete and partial cleanup of hazardous waste sites in the Clark Fork Basin. In the present report, I will provide a peer review of their study. For convenience, I will refer to this study as the "Clark Fork study." The Clark Fork study used contingent valuation, a method that estimates values from survey data. For convenience, I will refer to contingent valuation as "CV."

In addition to the body of my peer review, this report contains two appendices composed of draft papers that provide background material about my approach to evaluating the validity of CV studies. I will refer to these appendices as needed for elaboration on the principles that I will be applying throughout the review. In addition, at the very end of the report, I have included as Addendum 1 my responses to criticisms of an earlier draft of this review by experts of ARCO, which is the defendant in the natural resource damage case relating to the Clark Fork sites.

My peer review will focus on the validity of the Clark Fork study.[1] The term "validity" as used here is synonymous with what others, including the NOAA Panel (U.S. Department of Commerce 1993), have termed "reliability." Either term refers to the accuracy of the results. The issue is whether the results from the Clark Fork study are sufficiently valid to be used in

[1]The boundaries of my review need to be explicitly stated. I have not reviewed the completed questionnaires or the data files I have taken at face value the results of the statistical analyses performed as they are described in the report. I have not attempted to replicate the statistical procedures followed in arriving at those results. Also, I have not reviewed the various reports covering injuries at the Clark Fork sites or attempted to compare the injuries described in the contingent valuation surveys with the injuries described in those reports. Finally, my review stops with the estimated values for partial and complete cleanup at the Clark Fork sites and will not include procedures followed in the estimation of damages over time.

estimating the damages to Clark Fork Basin resources from releases of hazardous substances at the sites in question.

It will help if the definition of economic value is reviewed at the outset. Economic theorists think in terms of the economic well-being or "utility" enjoyed by individuals. Individuals enjoy utility to the extent that their preferences are satisfied. In economic theory, the preferences over alternative situations are translated into monetary economic values using the concept of indifference. Suppose, for example, that a consumer prefers Situation A to an alternative, Situation B. Situation B might involve a higher price than Situation A for a commodity that the consumer enjoys or the pollution of some environment that the consumer has an interest in. In either case, one measure of the economic value lost if the consumer were forced to accept Situation B rather than Situation A would be an amount of compensation sufficient to make her indifferent about the change from Situation A to Situation B once she received the compensation. This is the willingness to accept compensation or WTA measure of value. In this example, WTA is defined in a way that gives an economic interpretation to the legal concept of paying compensation to make the consumer "whole" after she has suffered a loss. A second measure of value would be the amount of money that the consumer would be willing to pay to stay at Situation A. This is the willingness to pay or WTP concept of value. The concept of indifference is central to both WTA and WTP. *Economic value* is defined as the amount of monetary compensation paid or received that would leave the consumer indifferent between Situations A and Situation B. WTP and WTA will be termed "indifference-producing" amounts of money. In the terminology of measurement theory, the concepts of WTP and WTA play the role of "true values."

It is important to recall the underlying definitions of economic value when thinking about empirical estimation of these values. Whether some sort of "revealed preference" method such as market demand estimation or some sort of "stated preference" method such as CV is used to estimate economic values, the goal is to measure indifference-producing amounts of money. This makes assessing the accuracy of any valuation method difficult. The reason is that indifference is a mental state and as such cannot be observed directly. If indifference cannot be observed, neither can theoretical WTP nor WTA Thus, it is impossible in principle to observe the true values of WTP or WTA and use the results as a standard against which to assess the accuracy of any method of valuation. This is true whether the accuracy of a revealed preference method or stated preference method is being considered. Rather than comparing estimated values to true values straight away, less direct methods of assessing accuracy must be sought.

This is not an unusual problem for the social sciences. Such concepts as intelligence and proficiency in mathematics, for example, are as unobservable as WTP and WTA. IQ tests and math exams provide evidence about intelligence and math proficiency, respectively. In a similar vein, market demand analyses, CV studies, and other methods are used to estimate economic values. In all these cases, the question becomes, How good (how valid? how reliable?) are the observed values as indicators of the unobservable true values?

Three mutually reinforcing approaches have evolved in the social sciences to address this question: content validity, construct validity, and criterion validity, the "Three Cs" so to speak. Each of these approaches to validity assessment can be applied to CV.

Content validity assessment involves examining the procedures that were followed in designing and executing the CV survey and analyzing the results. Poor procedures may lead to inaccuracy. Sound procedures should increase the likelihood that observed values are valid.

Construct validity assessment involves testing theory-based hypotheses about the relationships between estimated values and other variables. Construct validity assessment asks whether or not the relationships between contingent values and other variables, including other contingent values, are consistent with principles from the theory of value. Construct validity assessment may involve either convergent validity tests or theoretical validity tests. Convergent validity tests usually employ comparisons of contingent values with values derived from other nonmarket valuation procedures such as values from hedonic price or travel cost demand studies. Theoretical validity involves testing theoretically motivated hypotheses about relationships between responses to CV questions and other variables. So called "valuation equations" are often estimated using multiple regression techniques. Valuation equations use some value measure as the dependent variable; independent variables are chosen based on theory and intuition. For example, valuation equations are often used to investigate whether values are positively related to income. One would expect that respondents with higher incomes would, other things being equal, be willing to pay more for improvements in their circumstances and more to avoid losses.

Another example of a theoretical validity test is the a so-called "scope test." Scope tests involve comparisons of contingent values for two or more levels of provision of environmental resources. If respondents prefer one level to the other then they should be willing to pay more for the higher level. If they fail to do so, then theoretical expectations have not been satisfied and the validity of the CV study is called into question. Passing construct validity tests adds support for interpreting contingent values as valid estimates of true values. Failures of such tests raise doubts about such an interpretation.

Criterion validity assessment involves comparing CV measures to other value measures that are arguably closer to the true value. Although true values cannot be observed directly, if

opportunities are available to estimate true values in ways that have high scientific credibility, then such estimates can serve as standards for comparison, or criteria, for evaluating the validity of CV. Progress in criterion validity research depends on finding situations where the same good, service, or environmental amenity can be valued using CV and some other method that is generally accepted to be at least as accurate. CV involves transactions (broadly defined to include activities like voting in referenda on issues of service provision and taxation) that are fundamentally hypothetical. Most economists would agree that real transactions for the same good, service, or amenity should, other things being equal, yield values that are at least as valid as contingent values.[2] Thus, in criterion validity research, results from real transactions of some sort generally serve as the standard against which to evaluate the validity of contingent values.

A more complete exposition of the problem of CV validity assessment and how each of these approaches addresses that problem is provided in Appendix A of this review. This is a paper by Bishop, Champ, Brown, and McCollum presented in the summer of 1994 at an international conference on CV. Proceedings of that conference will soon be published as a book by Kluwer Academic Publishers. The bulk of this report will assess the content, construct, and, to the extent possible, the criterion validity of the Clark Fork study. First, however, a broader question will be raised. It would make little sense to delve into the detailed procedures in the Clark Fork study if CV itself has been shown to be invalid. Thus, I will first summarize current thinking on the overall validity of the CV method.

[2]There is a subtle but fundamental point here that is often overlooked. Economists sometimes forget that market values are *not* true values but only estimates of true values. This means that the measurement issues that arise in market valuation are basically the same as those that arise for contingent valuation, since both attempt to measure unobservable true values based on observable phenomena. Though the validity criteria themselves would need to be adapted somewhat to fit the special characteristics of market values, the basic principles embodied in the Three Cs are just as applicable to market values as they are to contingent values. The credibility of market values in the eyes of economists is evidently based, at least implicitly, on the presumed content validity of data on market transactions.

OVERALL VALIDITY OF THE CONTINGENT VALUATION METHOD

CV has evolved over the past three decades. As with any new approach in an established discipline like economics, CV generated a healthy debate among researchers that fostered its continued development and refinement. The Exxon Valdez oil spill turned out to be a watershed event in the evolution of the method. Fearing that the use of CV to assess damages would result in a large claim, Exxon employed some of the nation's leading economists to help prepare its defense. This has resulted in a raging national debate over the method. This controversy is no doubt confusing to outsiders, including business people, government officials, attorneys, and judges. Some of the most vehement opponents of CV are widely respected econometricians from some America's best universities. Other economists and econometricians, including two Nobel Laureates in Economics, believe that the technique is sufficiently valid to be useful in policy analysis and litigation.

Much of the most severe criticism is found in various chapters of the book edited by Hausman (1993). Further discussion of the critique may be found in papers by Diamond and Hausman (1994) and McFadden (1994) and the monograph by Desvousges et al (1992). I will attempt only a sketch of the dimensions of the debate here.

Those most critical of CV begin from the standard presupposition that only revealed preference data--and most notably data based on actual purchases and sales in markets--hold reliable information about economic values. Many economists are dubious about the credibility of verbal reports about economic preferences. That verbal reports may be untrustworthy goes back at least to Samuelson's (1954) classic theoretical article on public goods. Samuelson argued that economic agents would have incentives to give "false signals" (Samuelson 1954, p. 388) about

their preferences for goods and services consumed collectively in order to free ride while others pay the costs. Many environmental amenities have public goods characteristics. Hence, by implication, consumers would have an incentive to give false signals on CV surveys involving environmental values. However, Samuelson's argument was completely theoretical. The critics of CV agree that ultimately empirical evidence should be consulted to determine whether, despite concerns like those raised by Samuelson, CV data provides valid information about economic preferences. However, they believe that there is a major impediment to empirical research on the problem.

The critics reason that while market valuation methods are subject to external validation, CV is not. By external validation, they evidently mean tests comparable to what I call criterion validity tests. CV studies are thought to include nonuse or passive use values--values associated with motives other than personal consumption or other use of environmental resources. Because nonuse values are not based on resource use or other consumption activities, people who hold nonuse values will not leave sufficient market or other behavior-based evidence of their values to form the basis for external validation of contingent value estimates.[3] Hence, the critics reason that CV will at best remain inferior to market methods.

The critics of CV do recognize the possibility of construct validity testing, or "internal validation" in their terminology, though they consider it less potent than external validation. That is, they do recognize that CV might gain some economic credibility if it could produce results consistent with prior expectations based on economic theory. In the terminology of this report, they propose that CV results be subjected to strict theoretical validity testing. They carried out

[3]This is not to say that there is no behavioral evidence. For example, charitable giving for preservation of far away rain forest and whales may reflect nonuse values, but the evidence would be difficult to use in valuation due to free riding on the part of many members of the population.

three CV studies--the study of wilderness preservation values first discussed in Diamond et al. (1992) and later used in McFadden (1994) and the bird protection and oil spill valuation studies reported in Desvousges et al. (1992)-- and subjected the results to internal validity tests. When their studies failed such tests, they conclude that CV cannot stand up even to internal validity tests. Hence, they argue that CV is a flawed approach to measuring economic values.

Proponents of CV have been less than convinced by these arguments. Supporting arguments for the method can be found in many places including Mitchell and Carson (1989) and Hanemann (1994). The NOAA Panel on Contingent Valuation, a panel of distinguished scholars co-chaired by Nobel Laureates in Economics Kenneth Arrow and Robert Solow (U.S. Department of Commerce, 1993), concluded that well-done CV studies are sufficiently reliable to serve as a starting point for assessing natural resource damages. I will not attempt to do justice to all the arguments of those who support CV, but will attempt to summarize my own assessment.

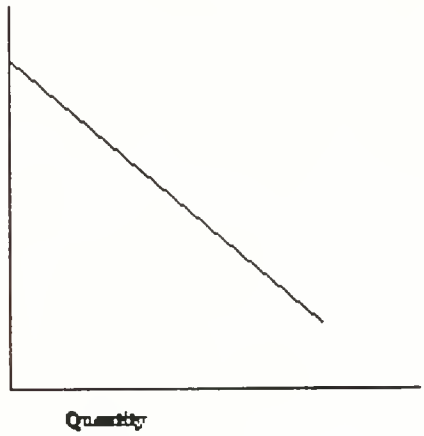

Figure 1

From the perspective developed in Appendix A, CV and revealed preference methods have more in common than the critics are admitting. To explain this point, let us consider the so-called "welfare triangle" that serves as the basis for valuation in market value studies as illustrated in Figure 1. This Figure shows a standard demand function for an individual consumer as portrayed in economic textbooks. It indicates the various quantities that this

consumer would purchase, symbolized by Q, at various prices, symbolized by P. The current market price is P' and this leads to consumption of Q'. P" symbolizes the price at which this consumer would no longer purchase the good. In everyday speech, this consumer could no longer "afford" to buy the commodity in question if its prices were P" or higher.[4] The welfare triangle is the area of "consumer surplus" under an individual's demand function and above the price line, the area P"A P' in the Figure.

Market valuation involves efforts to estimate demand functions in the real world. Then, welfare triangles or portions of them are interpreted as estimates of the economic values of market commodities. For example, suppose that the price of the commodity increases from P' to P" The consumer is worse off because she can no longer afford this good. What is the value of her loss? If she behaves like economic theory predicts she will, her consumer surplus will represent enough compensation to make her as well of as she would have been had the price not changed. Stated differently, the welfare triangle is an estimate of one of those indifference producing amounts of money that I discussed in the introduction. It is <u>an estimate</u> of an amount of money that would be just sufficient to make her indifferent between (1) having the opportunity to consume the market good at P' and (2) having the price of the good be P" but receiving the compensation. The question to be asked in any empirical application is, how good an estimate is it? We economists do not really know whether her observed consumer surplus would leave her indifferent or not. The gap between observable consumer surplus and the unobservable mental state of indifference is bridged with theory. ***If*** she behaves in a manner exactly like economic theory's consumer, then her demand function will be positioned so that the consumer surplus will

[4]In order to avoid certain technical issues that are not relevant to the discussion here, I assume that the so-called "income effect" associated with this demand function is zero.

measure her economic value for the item in question. Otherwise, her consumer surplus will not measure her economic value.

Therefore, the problem of establishing the validity of market estimates of value and CV estimates of value is fundamentally the same. In market valuation, we observe welfare triangles and infer indifference-producing values. In CV, we observe questionnaire responses and infer indifference producing values. Judging the validity of the inferences from observable to unobservable is the problem in both cases. Validity of market values, like the validity of contingent values, must be evaluated indirectly based on strategies that I have called content validity, construct validity, and criterion validity.

Given the long established position of market-based values, it is instructive to ask about the case that can be made for their validity. Though economists have not used the Three Cs explicitly, the principles that they embody are clearly present in how economists view market values. Content validity assessment comes in as economists examine the quality of the market data to be used in their analyses. One of the failings of economics as currently practiced is that analysts do not go very deeply into the matter. Rather, it seems to me to be fair to say that, within economics, market values are presumed to have high content validity unless there is a strong and obvious case to the contrary. That is to say, economists normally take data from market transactions at face value. Nothing much has been done to try to develop criteria for when market transactions are satisfactory indicators of value and when they are not. Dealing with this issue is beyond the scope of this report. I simply state it as an observation.

The critics of CV come from a market valuation tradition. As noted above, they have pointed to both external and internal validity tests. These tests are more or less synonymous with the criterion and construct validity tests of measurement theory. The external validity tests they

have in mind (Diamond and Hausman 1994) are tests of "predictive validity," a form of criterion validity. Demand equations would be estimated and then used to predict consumer behavior under specified conditions. The predictions could then be compared to actual behavior to evaluate the performance of the estimates. This is, of course, a sensible direction for research to take. To the extent that estimated demand models predict accurately, confidence in them is enhanced and to the extent that they do not, then it's back to the econometric drawing boards. At the same time, however, the underlying issue of how closely market values approximate indifference-producing economic values remains largely unaddressed by this approach. Predictive validity is a necessary condition for accurate valuation, but it is not a sufficient condition. A model might predict actual behavior well, but if consumers are not behaving the way theory describes, then the link between observed demand and the indifference-producing values would not be present. Thus, I disagree with the critics of CV when they say that external validity is more important than internal validity. The Three Cs are three "legs of the validity stool," and construct validity is an essential leg.

It is thus to their credit that economists routinely apply construct (internal) validity tests in market valuation studies. This is done each time a researcher examines the coefficients in market demand estimates for sign, significance, and more complex relationships that are expected to exist based on economic theory. And, it is instructive to see how well studies of market demand fare when tested. I am not an expert on market demand estimation and have not made anything close to a full review of the extensive literature on that subject. Nevertheless, some understanding of the current state of the science can be drawn from the work of Paris et al. (1993), who reviewed econometric articles dealing with production and consumption in the agricultural sector published in the preceding 74 years of the American Journal of Agricultural Economics and other outlets for

scholarly work. The agricultural sector has been one of the most intensively studied parts of the U.S. economy and the successes and failures that have occurred there should be rather indicative of the overall status of mainstream economic measurement. A quotation provides the flavor of what they found in their review (Paris et al., 1993, p. 37):

> A rigorous test of traditional consumer theory consists in verifying whether commodity demand functions are homogenous of zero degree in income and prices, and whether income-compensated price derivatives constitute a symmetric negative semidefinite matrix. Our literature review has shown how difficult it is to find a complete and satisfactory test of these conditions. Despite these unfavorable results, theorists and practitioners continue to derive policy conclusions from a theory that appears to be largely refuted.

In other words, prior expectations based on theory are so frequently violated in empirical work on market demand that Paris et al. feel that the theory itself is questionable.

What conclusion do economists draw from such failures? Do they throw out market valuation as an invalid tool? The answer is that they do not. As Paris et al. (1993) point out, progress in economics comes through an iterative process of developing theories, applying them empirically, and then developing new theories based on the empirical failures of theoretical predecessors. Simply rejecting the current theory is not enough. Instead, we economists proceed as best we can with our current theories until we find improvements. The consensus among economists is that we have made enough progress to provide market values that are useful in policy analysis, litigation, and private uses such as demand forecasting. Nevertheless, economics has a long way to go as a science. In the meantime, the more often our empirical market values match theory-driven prior expectations, the more confidence we should have in our results. At the same time, we should not be surprised when market valuation studies do not pass each and every construct validity test that is thrown at them. Neither our economic theory and nor our

empirical methods are yet perfected, and until more progress is made, imprecision in results and uncertainty about their validity will remain.

Now consider what all this means for CV. Given the fundamental similarities between the validity issues that arise for market values and contingent values, methodological consistency requires that parallel approaches be applied. Some economists question the content validity of CV data. I agree that serious pitfalls can be encountered in gathering CV data. Thus, at a minimum, validity requires that CV studies be done well. A high level of content validity is not a sufficient condition for valid contingent values, but it is a necessary condition. This is not some sort of weak standard, as Hausman (1995, p. 53), argues but one important leg of the three-legged stool on which validity rests. Criterion (external) validity testing is the second leg. Research on the validity of CV should stress comparisons between what people say they will pay for environmental amenities and what they will actually pay. Finally, construct validity testing should probe for strengths and weaknesses in the links between contingent values and theory-based prior expectations. The stronger such links are, the more confident is the researcher that the processes depicted by theory are in fact present in the data. I disagree with Hausman's conclusion that construct (internal) validity tests are weaker than criterion (external) validity tests (Hausman 1995, p. 53) Both are important. Construct validity testing is the most direct route to exploring the links between theory and empirical results.

Thus, there are two basic flaws in how the critics of CV have approached the validity issue. First, they have been inconsistent in how they treat CV studies and market studies. Flaws that, given the state of the science of economics, are accepted as unfortunate but unavoidable in market valuation studies are treated by the critics as fatal flaws in CV studies. I want CV studies to pass such tests even more than they do, but economics has not yet reached the level where such

tests can be consistently passed by either market or CV studies. Second, to attempt to draw sweeping conclusion about the validity of CV from a few studies, as the more vociferous critics of CV have done, cannot be justified scientifically. Critics of CV, who designed their own CV studies and tested their results, may have shown that their CV studies are invalid, but judgments about the overall validity of the method must be based on the preponderance of evidence across a full range of studies.

How has CV fared across a broad range of studies? As the critics themselves admit (Diamond and Hausman 1994), valuation equations often show relationships between expressions of WTP and socioeconomic variables such as income that are statistically significant and of the sign predicted by theory and common sense. A survey of CV studies also show that many have also passed scope tests (Carson 1994). These results support the view that values estimates from some CV studies, at least, are linked to economic preferences. If no such results were forthcoming, contingent values would have little economic credibility. The critics cannot have it both ways. If failures to pass construct validity tests are to be interpreted as negative evidence (as I agree they should), then passing such tests must be interpreted as positive evidence.

As I and my co-authors in the paper in Appendix A point out, criterion validity studies of CV have had mixed results. Sometimes CV has performed relatively well and sometimes it has not. The same thing could be said for studies that have compared contingent values of amenities to estimated values based on travel cost demand models, hedonic price methods, avoidance costs and other revealed preference techniques. This latter group of studies are probably best interpreted as convergent validity studies, as that term was developed in the introduction.[5] In a

[5] Even though applications of the travel cost and hedonic price methods and other such "indirect methods" involve revealed preference data, they have validity problems of their own and at least at this stage are not, in my view, sufficiently valid to warrant their being treated as criteria for criterion validity tests.

recent paper, Carson et al. (forthcoming) considered 83 separate studies that allowed for 616 comparisons of contingent values to revealed preference values for the same environmental services or other amenities. All the studies consulted involved WTP. Various comparisons of the contingent values and revealed preference values showed a rather close relationship. For example, the full 616 comparisons indicated that the ratio of contingent values to revealed preference values averaged 0.89 with a 95 percent confidence interval of [0.81-0.96] and a median of 0.75. The Spearman rank correlation coefficient for contingent values and associated revealed preference values was 0.78.

These results show that, across a wide range of studies, contingent values corresponded quite well to revealed preference values. Hence the overall validity of the CV method is supported at least up to a point. Admittedly, the comparisons of Carson et al. involved only studies where revealed preference estimates were possible. Nonuse values were not included. Admittedly, nonuse value studies do represent special challenges to CV, since the information communication requirements become much larger when dealing with members of the general public who have very limited or no past experience with the resources at issue. Still, after seeing the kind of performance record for CV documented by Carson et al., it is quite a leap to argue *a priori* (i.e., without the benefit of any empirical evidence that once the domain of nonuse values is entered, CV should have no credibility at all. It is much more plausible to expect respondents, who have been reasonably well informed through carefully prepared survey material about how environmental amenities will change, to do well enough in a CV exercise to provide useful information about their economic values for changes in environmental and other amenities.

Thus, though the debate continues, many economists have concluded that CV studies, if well done, are sufficiently accurate to be used in damage assessments and policy analyses. That

was certainly the conclusion of the NOAA Panel on Contingent Valuation (U.S. Department of Commerce, 1993, p. 4610):

> It has been argued in the literature and in comments addressed to the Panel that the results of CV studies are variable, sensitive to details of the survey instrument used, and vulnerable to upward bias. These arguments are plausible. However, some antagonists of the CV approach go so far as to suggest that there can be no useful information content to CV results. The Panel is unpersuaded by these extreme arguments.

The NOAA Panel stated many guidelines for CV studies. While not necessarily endorsing all these guidelines, I agree with the need for high standards and agree that if high standards are met, the NOAA Panel was correct to state (p. 4610):

> CV studies convey useful information. We think it is fair to describe such information as reliable by the standards that seem to be implicit in similar contexts, like market analysis for new and innovative products and the assessment of other damages normally allowed in court proceedings . Thus, the Panel concludes that CV studies can produce estimates reliable enough to be the starting point of a judicial process of damage assessment, including lost passive-use values.

Against this backdrop, the task of my report is to consider the strengths and weaknesses of the Clark Fork study. Does the Clark Fork study meet relatively high standards or does it have major, potentially fatal, flaws? I will address this question primarily from the perspectives of content and construct validity. In the next section, I will apply a content validity rating form for CV studies to the Clark Fork study. This form has evolved out of my research over the last year and a half and represents my attempt to distill the best ideas from several examinations of content validity issues including the NOAA Panel Report, Mitchell and Carson (1989), Hanemann (1994), and Fischoff and Furby (1988). The details of the rating form and the rationale for its various parts are explained in the paper in Appendix B. Next, I will turn to the construct validity of the Clark Fork study. For completeness and to avoid confusion, I will also consider the insights that are possible based on the criterion validity studies in the current literature. For reasons that will

be more fully addressed at that time, criterion validity studies are of limited usefulness in evaluating individual CV studies. Extrapolating from the results of simulated markets and other criterion validity studies to a field application of the CV like the Clark Fork study is a tricky business, because any attempt to do so is immediately clouded by methodological dissimilarities. This is a theme to which I will return later on.

CONTENT VALIDITY ASSESSMENT

Students of CV often conclude that one CV study is better than another. In part, such a judgment would be based on an examination of how the studies were designed and executed. In an effort to make the criteria that are applied in formulating such judgments more explicit and systematic, Bishop and McCollum developed a set of criteria that should be applied in assessing the content validity of CV studies. These criteria are stated in the form of questions that appear as the Content Validity Rating Form presented in Table 1 of our paper. My coauthor and I are proposing that reviewers of CV studies express their answers to the questions verbally and, for the first 12 questions about the details of study procedures, in terms of a numerical score. The scores are to express the extent to which the study under review meets the criteria in each case. As I evaluate the Clark Fork study and eventually two other studies as well, I will first present my answer to each question and then assign a score.

(1) Was the theoretical true value clearly and correctly defined?

The theoretical basis for this study is provided in Section 1.2 of the report and in formal definitions of the values estimated, which appear on pages 5-23. Equation 5-2 defines willingness

to pay as the income change required to exactly offset the utility gain from an improvement in the environmental amenities of the Clark Fork NPL sites. This definition of value causes no particular difficulty for cases in which a certain and rather immediate change in the status of the environmental resources is to be valued. In the Clark Fork study, the formal definition of value does not explicitly account for the uncertainty or the timing of any changes in the environmental attributes if a cleanup is undertaken.

The issue of uncertainty may not play a critical role in this study. As part of the scenario, respondents were provided with descriptions of what would be accomplished under the various proposed interventions. These effects were simply described without any sense of uncertainty attached to them. Perhaps of more concern is the issue of timing. Willingness-to-pay is expressed in terms of an annual payment during each of the next ten years. The connection between the time path of costs (the annual payments for ten years) and the time path of benefits of the intervention is not dealt with theoretically.

Another theoretical concern focuses on the definition of damages. Complete cleanup is used as the baseline for calculating damages and damages are taken as the "residual value," the difference between the value of complete cleanup and the value of partial cleanup. The true baseline condition of the resources is the condition that would have prevailed had the injuries never occurred. Since complete cleanup takes time, one would expect its value to be less than the value the resources would have generated had they never been injured (the theoretical value of "true damages"). Therefore, residual value as defined in the Clark Fork study (the value of complete cleanup minus the value of partial cleanup) will be less than true damages (the value of uninjured resources minus the value of partial cleanup), all else equal. The report does not consider this issue nor does it point out that residual value underestimates true damages, all else

equal. In this way a downward bias was introduced into the estimated damages and the study would have been stronger had it clearly analyzed this issue in theory and called attention to this bias.

One other theoretical aspect of the damage calculation needs to be considered: Apportionment was carried out to allocate estimated total values among various resource categories. This apportionment was accomplished using respondents' reports about the relative percent of total willingness to pay attributable to each of the various resource categories. Such apportionment questions are not only difficult for respondents to deal with, but also pose some theoretical difficulties. The value of any one of the components is theoretically dependent on the levels of the others.

In view of these theoretical concerns, I assigned 3 out of the possible 5 points on this item. However, I would quickly add that, while I might have some theoretical qualms about apportionment, I am not aware of any methods that are both practical and theoretically sound. There is precedent in the existing literature for what was done to apportion values in the Clark Fork study. Although questions arise from a theoretical perspective, the authors appear to have applied a reasonable practical expedient to sidestep a theoretically insoluble problem. It is not clear that apportionment led to either an upward or downward bias in the final value estimates, other things being equal.

(2) Were the environmental attributes relevant to potential subjects fully identified?

The Clark Fork study devoted substantial effort to qualitative research, including both verbal protocols and self-administered pretests. Although the researchers used different terms,

they deliberately set out at the beginning of their study to identify participant-relevant attributes of the resources in question. They speak (p. 2-16) in terms of "a process of acquiring a substantial amount of potentially relevant information about natural resource injuries, both in general and at the Clark Fork sites in specific, and formally paring down the information to retain the most critical information that respondents use to evaluate the sites and to form values." In the process of attempting to pare down the amount of information presented in the survey instrument, they presumably would have learned a great deal about which resources respondents consider relevant.

Nevertheless, some respondent-relevant attributes might not have been fully explored. Of particular concern was the treatment of wildlife and its habitats. Respondents were informed of a loss of 20 square miles of habitat used by elk, deer, otter, pine martins, great blue herons, grouse, birds of prey including redtail hawks, and common songbirds. The survey did not provide information about whether this has resulted in any reductions in wildlife populations that might utilize this habitat nor did it place the loss of wildlife (if any) into any sort of context.

It is difficult to know how much to make of this shortcoming. It might have been established during the qualitative research that habitats, rather than wildlife populations, were the critical piece of information required by respondents. Similarly, respondents who cared specifically about wildlife populations might have possessed sufficient knowledge about the consequences of the loss of habitats for the species involved. In either instance, the lack of detail about wildlife populations would not be a significant issue.

Furthermore, the study did better at explaining the extent of the injuries in other respects. As one example, the final survey instrument stated that trout have been eliminated from Silver Bow Creek and that trout populations in the Clark Fork River between Warm Springs Ponds and Milltown Reservoir have been reduced to about one-fourth of the population that would be

present if the contamination had not occurred. A bit more might have been said about how productive these fish habitats would have been under baseline conditions. In a similar vein, on the subject of groundwater, the survey presented some information. The questionnaire indicated that the water supply for the city of Butte was contaminated and needed to be replaced. It specifically mentioned that the residents of Milltown have also had to replace their water supply because of groundwater contamination. It explained how much water an acre foot is both in physical dimensions and in terms of how much water an average household uses. However, more might have been said about the number of humans who are affected by the contamination of groundwater or will be affected in the future. Such potential deficiencies can be mentioned, but it is difficult to say what if any biases were introduced because of them.

I assigned 8 points out of a possible 10 to this aspect. The qualitative research performed in designing this study supports the use of the information that was provided (see Table C-1, for example). However, some aspects, especially effects on wildlife populations, may have needed more attention. It is not at all clear whether these potential shortcomings biased the results, and if so, in what direction. Still, such loose ends reduce content validity.

(3) Were the potential effects of the intervention on environmental attributes and other economic parameters adequately documented and communicated?

The report on the Clark Fork study points out that the researchers worked with those assessing the injury to document how attributes were affected by the releases. This effort should have been adequate to document such effects.

However, regarding the communication of the effects, the instrument required that respondents read and absorb large blocks of information. This could have left some respondents less than well informed simply because they lacked the ability to absorb the information provided

and therefore could not use it in arriving at their values. It is possible that paying respondents $20 to complete the survey led respondents to work harder than they otherwise would have to read and digest the material. Still, the length and complexity of the survey left me with enough concerns about effectiveness in communication to warrant assigning only 6 out of 10 points on this item.

(4) Were respondents aware of their budget constraints and of the existence and status of environmental and other substitutes?

Individuals were not explicitly reminded of their budget constraints in the survey instruments. However, I do not view this as a flaw. The discussion of additional cleanup sites (for example, the NPL sites not in the Clark Fork Basin) and the discussion of other non-environmental programs must have served to heighten respondents' awareness of the fact that there are many competing demands for their limited resources. Likewise, the second question in the survey asked respondents to rank the relative importance of dealing with a series of problems. While the researchers indicated that this question was inserted at the beginning of the survey in an attempt to " . . . diffuse any importance bias to a specific topic that may result by receiving a survey on a specific topic . . . ," the question may also have served to remind respondents about other social issues that could require additional resources if addressed.

In my own work, I have found that respondents often spontaneously identify budget constraints, or income, when asked how they determined their responses to a willingness-to-pay question. In the Clark Fork study, respondents were encouraged to provide written comments and nearly 75 percent of them did so. The report stated (p. 5-18), "The most prevalent type of

comment (258 made) was simply an affirmation of the WTP response. These comments frequently referenced the respondent's income level as a determinant of WTP."

Turning to substitutes, it is helpful to remember the main motivation for being concerned about substitutes in damage assessments. The fear on the part of the NOAA Panel, as well as others who have critically reviewed the CV method, was that respondents would get the impression that the injuries are more widespread and injurious than they are in reality. For example, in evaluating an oil spill, respondents might get the idea that the spill affected vast areas and killed a large proportion of the bird and other marine life of the region when this is, in fact, not the case. Media coverage, for instance, might lead to such a misimpressions. If so, respondents would need to be reminded that large areas of the coast and associated wildlife populations were unaffected. Thus the NOAA Panel was particularly concerned that respondents be informed about the extent of undamaged substitutes.

Was this a serious problem for the Clark Fork study? In my judgement, it was most likely not a problem. One reason is that the release took place in Montana and only Montana residents were included in the study. It is likely that most Montana residents are aware that their state abounds in pristine areas that have not been affected by release of toxic materials to any great extent. Second, the survey instrument itself contained a great deal of information that indicated that the geographic extent of the injuries was limited. The cover of the survey was a map that clearly conveyed the message that the problem was limited to one relatively small region in the southwestern corner of the state. A color map (Map A) was included in the visual materials that shows how only a small part of the region shown on the cover of the questionnaire was affected. It clearly indicates that percentage of the land area (including wildlife habitats) that was included within the affected sites and shows several unimpacted streams including two that would be

familiar to many respondents, the Blackfoot River and the Bitterroot River. Third, in addition to identifying some possible environmental substitutes, the survey instrument also asked survey respondents to rank the importance of several non-environmental problems facing society at large (e.g., improving education, reducing air pollution, bringing in new jobs). The positioning of this question prior to the valuation question may well have served to remind respondents of a broader range of social problems that might require substantial resources to solve.

Also, I would argue that it may be as important to emphasize damaged as well as undamaged substitutes. In this case, respondents may have needed to know the extent of the problem of contaminated sites within their state as much as they needed to know the extent of the undamaged areas. The questionnaire identified and briefly described all of the National Priority List (NPL) sites in the state of Montana and mentioned the total number of other sites at which contamination may have occurred. This helped respondents view the Clark Fork sites in the broader context of the total number of sites that might eventually require attention.

Considering all this, and especially given that the survey was administered only to Montana residents, my conclusion is that the Clark Fork study did a more than adequate job of dealing with budget constrains and substitutes and I awarded it 5 out of 5 points on this item.

(5) Was the context for valuation fully specified and incentive compatible?

The context for valuation in this study seems to have been generally adequate. Respondents were asked to indicate, using a payment card, the maximum amount they would be willing to pay each year for the next ten years. The payment would be used to finance the cleanup of the Clark Fork NPL sites. Respondents were informed that all members of society and

specifically all Montana residents would have to pay part of the cost of cleaning up these sites, and that private industry and agencies of the U.S. Government were already paying for part of the cleanup. Mentioning payments by entities who caused the contamination is of particular importance. Otherwise, respondents tend to report lower WTP values than they otherwise would or even reject the whole valuation exercise. This would be an undesirable bias to introduce because under these circumstances such respondents do not express their true values for the resources in question but protests about having to pay for cleanup when those responsible are not paying.

An unusual feature of the context in this study was the use of multiple payment vehicles. Just prior to the valuation question, survey respondents were asked to evaluate the acceptability of various payment vehicles that could be used to collect the money required for cleanup. Respondents were asked to assume that their stated willingness to pay would be collected using one or more of the vehicles identified as acceptable by them in this earlier question. This approach was taken in order to reduce protest bids emanating from undesirable payment vehicles and it may indeed have done so.

A second unusual feature of the context was that respondents were informed that, "If cleanup efforts cost less than people are willing to pay, the fees would be lowered so that everyone would pay only a share of what the cleanup actually costs." This feature of the context may also have served to reduce protest bids. Respondents sometimes express skepticism about whether money committed to environmental remediation will actually be spent on it. For example, during the verbal protocols in the Clark Fork study, one subject said, "Well, I think if they would assure me as a Montana resident that my money would only go to the cleanup of the Clark Fork Site and that somehow we could manage to keep the bureaucrats away from the

money, so that they don't hire their own assistants for their assistants, I would say that I wouldn't mind an increase in my state or property or gasoline tax . . . ."

Respondents may also sometimes try to calculate the amount of money that would be raised if a specified amount was collected from all citizens. A hint of this perspective was also revealed in the verbal protocols when one respondent said, "If the method was more on fishing and hunting licenses, I think $3 would add up to be a lot."

Potential biases associated with either of these points of view may have been reduced by specifying that fees would be lowered if the cost of the cleanup proved to be less than the money that was collected Thus, treatments of the payment vehicle and surplus funds enhanced the validity of the results.

On the other hand, some possible reservations about the context also come to mind. The timing of the environmental improvement was not explicitly stated in the CV scenario. Other reservations arise as well: While this study's rather unique approach to the choice of payment vehicle may have reduced payment vehicle bias, it may have introduced perverse incentives as well And, the use of a payment card as opposed to other possible CV question formats may have biased results downward. These concerns will be discussed in turn.

Economic theory would lead one to expect that timing of the cleanup could be important to the values that people express. For potential users, lack of information on timing could have been a particularly important ambiguity. Some respondents may have assumed that, since they were to pay over 10 years, cleanup would take 10 years, but there is no way to know whether this was the case for everyone. This ambiguity reduces the content validity of the study. To the extent that temporal ambiguity actually affected responses, it probably did so in a negative direction. It is hard to imagine plausible reasons why respondents would increase their values as a

result of feeling uncertain about when cleanup would be completed, unless they somehow assumed that cleanup would be completed more promptly than the researchers intended. I see nothing in the questionnaire or associated materials that would have caused respondents to misinterpret the survey in that way.

When CV researchers advocate contexts that are "incentive compatible" they have a specific set of theoretical ideas in mind. On the broadest level, some incentive to misstate preferences will be present in most CV exercises. Unless respondents fully believe that their responses will directly determine how much of the amenity they will receive and how much they will pay, an incentive will exist for them to misrepresent their true values. As soon as they recognize that their responses involve hypothetical commitments to pay, they will have an incentive to answer so as to enhance the chances of a favorable outcome or reduce the chances of an unfavorable outcome. Whether respondents believe the scenario is covered under Question 6 below. Here, a narrower set of incentives is being referred to. Incentive compatibility presupposes that respondents fully believe the CV scenario, and then asks whether they have every incentive to respond to the CV question in a manner consistent with their true values. For example, it is well established in the literature (see, for example, Hoehn and Randall (1987)) that a referendum format with single-bounded responses (i.e., where the respondent is asked whether or not she would vote in favor of a proposal given the cost to her specified in the question) is incentive compatible. That is, if she really believes that her "vote" in the survey will be counted in determining whether the proposal is adopted and that adoption will mean that she will have to pay the specified amount, then theory indicates that she will vote "yes" if her true WTP exceeds the specified amount and "no" otherwise. On the other hand, real world sealed-bid auctions where winning bidders pay their bids are known to be incentive incompatible. Theory indicates that

bidders in real auctions organized in this way will have an incentive to bid less than their true values for the item being offered in order to try to "get a good deal." To the extent that open-ended CV questions mimic sealed bid auctions, they might not be incentive compatible.

To examine whether the Clark Fork study's payment vehicle and payment card created any perverse incentives, let us assume that the respondents believed their answers would affect whether cleanup takes place and how much they would pay. Consider the payment vehicle first. As noted the respondents got to select their own, a rather novel approach to trying to avoid vehicle bias. There is the risk, however, that respondents might have tried to seek a strategic advantage by their choices of payment vehicles. For example, Survey Question 27 allowed them to choose highway tolls for Interstate Highway 90 in the affected area. Is it possible that substantial numbers of respondents chose this option and then answered the CV question (Survey Question 28) with a positive amount, knowing that they never travel this highway and thus could get out of paying what they had said they would pay.

However, both the wording of the two questions and the results associated with them make this an unlikely outcome. The first four responses to Survey Question 27 explicitly discouraged respondents from choosing payment vehicles they that they were not already using. For example, the second response (emphasis added) was "increase in waste disposal (trash collection) bills <u>you pay</u>." Respondents who do not pay for their own trash collection would be steered away from this response by the condition that is underscored.

In addition, it is clear from Table 5-12 of the January 1995 report that many respondents checked off categories that would require them to pay (68.5 percent chose increases in disposal fees and taxes on industry that would be passed to consumers in higher prices, 39.5 percent checked waste disposal bills they pay, 22.4 percent chose water bills they pay, and 21.7 percent

chose state taxes they pay (note that multiple responses were permitted so that these percentages total more than 100 percent). Furthermore, Survey Question 28 asked (emphasis in original), "What is **the most** that your household would be willing to pay each year for 10 years through the methods you selected in Q27 . . . ?" This wording leave little room for respondents to assume that their households could get out of paying by their having answered Question 27 strategically.

To the extent that respondents set up their answers to behave strategically by choosing vehicles that were not binding on them, one would think they would have bid relatively high amounts. Yet, as Table 5-13 shows, people who chose potentially non-binding payment vehicles in Question 27, agreed to pay very low amounts on average in Question 28. Thus, the explanation that the researchers proposed, namely that respondents used Question 27 to vent their protests at having to pay for fixing a problem created by others, seems much more likely than the alternative that Question 27's opportunity to chose non-binding payment vehicles set up opportunities to overbid in Question 28 for strategic reasons. If some respondents assumed that cleanup would go forward and perceived that answering Question 27 with a non-binding vehicle would mean that other would pay--an even more unlikely strategy in my opinion--the resulting bias would be in a negative direction.

Lastly, consider the incentive properties of payment cards. As noted above, CV questions with single-bounded referenda responses are thought to be incentive compatible and open-ended questions may not be. What is known about the incentive properties of payment cards? To my knowledge this question has not been addressed in terms of formal theory. This leaves some doubts about the incentive compatibility of the approach taken in the Clark Fork study, but I tend to think that the payment card as it was set up there has satisfactory incentive properties. In particular, the provisions (in the preamble to Q28) that all money would be spent to clean up the

Clark Fork sites and that "If cleanup efforts cost less than people are willing to pay then fees would be lowered so that everyone would pay only a share . . ." were probably sufficient to show incentive compatibility in a formal analysis.

My conclusion then is that, except for neglecting the timing of cleanup, the context for this study was adequate. I therefore awarded it 7 out of 10 points for this item.

(6) Did survey participants accept the scenario? Did they believe the scenario?

Scenario acceptance is a difficult issue to assess. The investigators clearly made an effort to design a scenario that was plausible to the respondents. Allowing them to choose an acceptable payment vehicle should also have increased the acceptance of the scenario. One bit of qualitative evidence of scenario acceptance is the result that only 46 out of the total of 841 responses (5.5 percent) were clearly identified as protest zeroes and only 21 (2.5 percent) were identified as outliers. This is evidence that a large percentage of the respondents accepted the scenario.

On the other hand, the survey asked if respondents felt they were personally responsible for paying part of the cost of cleanup. More than half of the respondents in the cleaned data set (Table 5-22) felt little or no responsibility for paying for cleanup (i.e., they chose 1 or 2 on a seven-point scale, where 1 represented "not at all responsible"). About 30 percent of respondents (p. 5-18 of study report) made comments protesting having to pay for cleanup. This is symptomatic of scenario rejection at some level. It would have led to a downward bias in the final value estimates, except that the authors controlled statistically for the resulting bias in arriving at their final value estimates.

Moving beyond acceptance, respondents would be said to have "believed" the scenario if they believed their responses would actually affect both the flows of environmental services from the Clark Fork sites and what they would actually pay. At least half of the respondents claimed to be familiar with one or more of the Clark Fork sites. The investigators built on this awareness by explaining in the survey that ". . . to make decisions about cleanup programs that could cost you money, it is important to know how much it is worth to you to clean up the Clark Fork River basin." The cover letter stated that study results would be considered when future decisions about potential cleanups were made. These steps helped to make the scenario believable.

On the other hand, the survey instrument did not make clear how the survey responses would be translated into decisions about Clark Fork resources. Furthermore, the survey did not contain any questions asking if the respondents felt that actual decisions regarding cleanups would be based on the results of the survey. The flexibility of the payment vehicle, which had the advantages already noted, had the disadvantage of making the whole exercise seem more hypothetical. It is difficult to determine whether survey respondents really believed that responses to the survey would affect decisions about the level of cleanup and their own expenditure patterns.

As noted above, disbelief in a CV scenario increases the incentives to respond strategically. These incentives have been a cause of concern for the thirty or so years that CV studies have been conducted. Nevertheless, there is precious little evidence that these incentives have been sufficiently strong to seriously bias value estimates. The incentives themselves are more complicated than a superficial examination of the problem would indicate. Mitchell and Carson (1989, pp. 158-165) take the problem apart in some depth. I will give a somewhat more limited view using the Clark Fork study as an example. Suppose that I am a respondent, that I do

not believe that how I answer will affect how much I will actually have to pay, and that I wish to respond in my own best interest Suppose further that I hold some positive value for fixing up the site (i e., my true value exceeds zero). Now, in order to decide on a strategy, I need to know how much it will cost me if the sites are cleaned up. If I think the cost to me will be less than my true value then my strategic response is to reveal a very high value to encourage clean up. If I think the cost to me will exceed my true value I will submit a very low, possibly zero bid. Now suppose that everyone is like me except that we have different true values and different expectations about how much cleanup will cost us. Then, a very plausible result would be a wide "bi-modal" distribution of values. That is, bids will be concentrated on the low and high ends of the spectrum. Of course, other possible sets of assumptions exist about the incentives that individual respondents will face depending on what they believe about how their responses will affect what will happen and what they will pay. If I do not believe my response will affect what happens at the Clark Fork sites, then my optimal strategy will be to refuse to take the time and trouble to respond at all. If I decide that the survey will not affect whether the site is cleaned up and that my response to the valuation question will affect how much I actually have to pay (an unlikely view in this case) then I will bid low or zero. The end result from all the possibilities would be likely be wide bi-modal distribution of strategic responses concentrated at relatively high and relatively low values. While many studies, including the Clark Fork study, show a high frequency of zero values, studies with large numbers of relatively high values are rare. The distribution of values shown in Table 5-6 of the January 1995 report from the Clark Fork study looks quite normal with large groups of responses at zero and in the middle of the range of responses Except for some protest zeros, there is little evidence of strategic responses here.

The question that led to this argument had to do with whether respondents, if they perceived (correctly) that their responses to the survey would not affect what they actually pay, might respond strategically encourage cleanup while others paid the tab. Because of the dearth of evidence of strategic responses in CV studies, content validity does not require that a large majority of respondent actually believe that their responses will affect both the future availability of amenities in question and what they will really pay.

Given that the Clark Fork study did a good job of investigating and correcting for respondent non-acceptance but did not achieve a high degree of believability, I assigned 8 out of a possible 10 points for Question 6. Basically, I am holding back 2 points for studies that go the extra mile of believability.

(7) How adequate and complete were survey questions other than those designed to elicit values?

I found the non-valuation questions to be adequate. Many questions were asked that could be used in assessing the construct validity of the study. These included questions related to potential uses of the affected resources (Questions 23, 24 and 25), importance of cleaning up NPL sites in Montana (Questions 15 through 22), and the demographic questions (Questions 35 through 44). Respondents were provided with an opportunity to provide written comments about their reactions to the valuation question.

A total of 8 out of 10 points was assigned here. The full 10 points were not assigned because other questions might have been asked. For example, while the study did ask about environmentally related behavior (recycling, organization membership, etc.) I would have been tempted to add some standard questions focusing on environmental attitudes.

(8) Was the survey mode appropriate?

Contrary to the conclusions of the NOAA Panel, I believe that mail surveys can be constructed to adequately carry out a CV study for natural resource damage assessment. Nevertheless, in this particular case, personal interviews might have been a better choice. A large amount of information was provided to respondents, perhaps approaching or even exceeding the limits of what can be reasonably accomplished using a mail survey. This could have been a partial contributor to the apparent failure of a full between-sample scope test, for example. (I will return to the scope issue in addressing the construct validity later on.) Personal interviews might have resulted in a higher degree of confidence that information in the survey instrument was successfully communicated to survey respondents. I should note in passing that, although

personal interviews might have been preferable from a scientific standpoint in this case, adequate personal interviews would have cost hundreds of dollars more per survey. It is not surprising that I and others do mostly mail surveys in CV studies. Whether the potentially improved communication would have been worth the extra cost is an open question.

Using an incentive as high as $20 has not, to my knowledge, been done previously in a CV survey. In addition to its likely contribution toward a higher response rate, it is plausible that the $20 encouraged many respondents to read the material and respond carefully, counterbalancing some of the potential adverse effects of using a mail survey. What other effects the incentive might have had are not known. From a theoretical perspective, $20 is small relative to income levels of respondents and the amount paid to respondents was independent of the values they expressed in the CV exercise. Consequently, one would hypothesize a negligible effect of the incentive on results, but to my knowledge this hypothesis has not been tested. Given the high costs of personal interviews, large incentives for mail responses may be a cost-effective substitute and should be researched.

Concerns about potential problems exacerbated by the mail survey are sufficiently strong to assign only 6 points out of 10 for survey mode.

(9) Were qualitative research procedures, pretests, and pilots sufficient to find and remedy identifiable flaws in the instrument and associated materials?

The investigators carried out three separate rounds of qualitative research, which included one set of verbal protocols and two pretests. Participants in the pretests were asked to fill out early drafts of survey instruments. After filling the survey out, they were engaged in a debriefing process.

Determining whether this was a sufficient amount of preliminary research is difficult. At the conclusion of the second pretest, the preliminary results appeared very favorable in terms of relative values of mean willingness to pay. In particular, the prospects for a cross-sample scope test appeared favorable. Furthermore, estimates of mean willingness to pay were nearly identical regardless of whether the complete cleanup was valued first or second. It might have been prudent to perform a more extensive pilot test that would have involved larger sample sizes to determine whether the results obtained in pretests, which had relatively small sample sizes per survey version, would be replicated in a true mail survey format. Furthermore, a more extensive pilot test could have revealed whether responses from the pretests were artifacts of the locations (Helena and Missoula) in which they were conducted. However, it is only fair to add that, while some of my concern has focused on the apparent failure of this study in a cross-sample scope test, the idea of a scope test had not been formally defined at the time this research was designed.

I assigned this study 4 points out of a possible 5 for this item.

(10) Given study objectives, how adequate were procedures employed to choose study subjects, assign them to treatments (if applicable), and encourage high response rates?

The sample used for this study was obtained from a highly regarded commercial vendor, Survey Sampling Inc. The sample frame was Montana telephone listings. Samples of this type are always subject to problems of non-coverage because some households have no telephone and others have unlisted numbers. In the case of the Clark Fork study, non-coverage is not a major problem for two reasons. First, the study reported that approximately 83 percent of Montana households have listed telephones. There are certainly many precedents in the CV literature for sampling coverage at this level. I would judge the level of coverage to be acceptable given that

coverage at higher levels of accuracy would have been difficult and expensive to obtain. Second, econometric procedures were used to adjust mean value estimates to account for both non-coverage and non-response.

The response rate in this study was relatively high (68.1 percent), particularly in light of the large amount of information respondents were expected to read and consider when answering the questions. Non-response bias does not appear to be an issue, particularly after applying the econometric procedure just mentioned.

Based on the conclusion that the study is basically sound with regard to sampling and survey procedures combined with minor concerns about sampling non-coverage, I assigned 8 points out of a possible 10 points.

(11) Was the econometric analysis adequate?

The econometric procedures applied were sound. The researchers were able to use standard regression techniques to define statistically significant valuation equations. The valuation equations included several variables that supported the construct validity of the study, as will be discussed shortly. Furthermore, the econometric analysis was used to carry out adjustments to willingness to pay to account for potential biases associated with non-coverage, non-response, and feelings of responsibility. The procedures used to identify protest zeroes and outliers made sense. While the procedures that were used for sorting zeroes and outliers would, all else equal, introduce a downward bias into the values for complete and partial cleanup, the impact on residual values was small.

I think that the econometric analysis done in this study was adequate, but did not include anything very original or innovative. Accordingly, I assigned this study 8 points out of a possible 10 for econometric procedures, reflecting that more analysis might have been done.

(12) How adequate are the written materials from the study?

Implementation procedures, response rates, analytical procedures, and results were all covered fairly well in the written report. I would have liked more details on some issues such as design procedures. I assigned the study 4 points out of 5 for reporting.

(13) TOTAL POINTS:

Table 1 summarizes the numerical scores on the first 12 items of the form and sums them up.

(14) Are there other concerns relating to the design and execution of the study that have not already been addressed?

I have only one other concern to raise. I would not have followed the procedure used in this study to address the embedding issue. I question whether embedding is a real problem in studies like this one. The most widely cited paper on embedding is Kahneman and Knetsch (1992), but I believe that that study has very serious flaws the make it an unreliable guide to the nature and extent of embedding problems in CV studies generally. Among its many problems, the Kahneman and Knetsch study did not identify and communicate the specific nature of the various interventions it attempted to value; the effects of the interventions were only vaguely described to respondents; the contexts for the CV exercises were typically poorly developed; and the survey

mode was telephone interviews. I tend to think that embedding does not have a very powerful influence in studies that are as specific in their scenarios as the Clark Fork study.

Table 1: Content Validity Scores for the Clark Fork Study

Questions (See Appendix B for further explanation; questions taken from Table 1 of Appendix B.)	Score/Total possible
(1) Was the theoretical true value clearly and correctly defined?	3/5
(2) Were the environmental attributes relevant to potential subjects fully identified?	8/10
(3) Were the potential effects of the intervention on environmental attributes and other economic parameters adequately documented and communicated?	6/10
(4) Were respondents aware of their budget constraints and of the existence and status of environmental and other substitutes?	5/5
(5) Was the context for valuation fully specified and incentive compatible?	7/10
(6) Did survey participants accept the scenario? Did they believe the scenario?	8/10
(7) How adequate and complete were survey questions other than those designed to elicit values?	8/10
(8) Was the survey mode appropriate?	6/10
(9) Were qualitative research procedures, pretests, and pilots sufficient to find and remedy identifiable flaws in the instrument and associated materials?	4/5

(10) Given study objectives, how adequate were procedures employed to choose study subjects, assign them to treatments (if applicable), and encourage high response rates?	8/10
(11) Was the econometric analysis adequate?	8/10
(12) How adequate are the written materials from the study?	4/5
(13) TOTAL POINTS:	75/100

If embedding is not really a problem, then Survey Question 30, asking respondents to "disembed" their previously stated values, is unnecessary. Using the responses to Survey Question 30 to reduce the estimated values of both cleanup options, as is done in the Clark Fork study, introduces a downward bias into the resulting value estimates. The presence or absence of embedding in responses to CV questions is an issue that is still being debated by researchers.

(15) Considering the issues raised in Questions 1 through 12, your total score as calculated for Question 13, and any additional issues raised under Question 14, how would you rate this study overall?

At this point, my rating form calls for a final qualitative rating ranging on a five-point scale beginning with "Excellent" and ending with "Unacceptable (Study Fatally Flawed)." In reviewing the individual scores, the Clark Fork study did quite well for the most part, earning 80 percent or more of total points on most items. More serious concerns arose regarding the adequacy of the theory; the choice of a mail survey mode; and the adequacy of the procedures to identify respondent-relevant attributes, to communicate the effects of the intervention, and to specify the context for valuation. Even where such concerns existed, however, scores equal to at least 60

percent of total points were earned. A score of 7 out of a possible 10 points was assigned to the acceptability and believability of the scenario, because the scenario might have been more believable, but more believable scenarios are not all that common. The total score of 75 out of a possible 100 points signifies a relatively strong study. I would assign it a rating of "Good" for content validity. It has many strong features but also some flaws that make an "Excellent" rating unjustifiable.

CONSTRUCT VALIDITY ASSESSMENT

As I have already explained, for contingent nonuse value studies, construct validity assessment involves statistical testing of theory-motivated hypotheses about relationships between answers to the valuation question and other variables. When a study passes construct validity tests this provides evidence that the processes depicted by economic theory are at work in the minds of respondents as they answer CV questions. This in turn supports interpreting responses as estimates of true values, as defined in theory.

Economic theory provides a great many possible hypotheses that could be tested. These range from the general observation that people with strong preferences for an environmental intervention should express higher willingness-to-pay values to the notion that higher incomes, all else equal, might be expected to result in higher willingness to pay. Such hypotheses can be tested using regression techniques in which willingness to pay is predicted by variables measuring reported behaviors, attitudes, income, and various socioeconomic characteristics. In the paper in Appendix A, we refer to such tests as *rudimentary tests*. We also consider more *advanced tests*, which normally involve comparisons of two or more contingent values. Scope tests are examples

of advanced tests. Scope tests investigate the degree to which willingness to pay is systematically related to the dimensions of the environmental intervention. One would expect, all else equal, that willingness to pay would be larger for an intervention that results in either higher quality or quantity of environmental attributes than for an intervention resulting in lower quality or quantity of environmental attributes.

Advanced tests are not limited to scope tests. For example, one could compare the value sum of the values for two environmental amenities, A and B, estimated using separate samples with the value A and B together, estimated using a third sample. This is the adding up test proposed by Diamond and Hausman (1994). Another type of construct validity test could be based on the concept of transitivity. If intervention X is valued more highly than intervention Y by one sample, and intervention Y is more highly valued than intervention Z by another, a third sample should place a higher value on X than on Z. Scope tests and other advanced tests are becoming increasingly important to the evaluation of CV studies.

The more construct validity tests that a study passes the better. However, failure to pass each and every test posed during an assessment should not be considered fatal. As we saw earlier, econometric studies utilizing market data often fail theoretically motivated tests, yet are considered useful. Economic theory has substantial limitations in predicting and explaining behavior. Econometric methods have their limitations as well. As I noted earlier, progress is made as these theoretical and empirical limitations are overcome. In the meantime, and despite imperfections, useful results are often obtained. This principle applies to CV studies as much as it applies to market studies. Construct validity is not a black-and-white issue, but a matter of degree.

To capture this principle, the paper included as Appendix A of this report suggests that studies be classified into a three-level hierarchy expressing increasing levels of construct validity. Quoting from that paper:

> At the lowest level would be studies that either have not included any construct validity tests or have failed to pass rudimentary tests. Such studies might typically have had low budgets and/or severe time constraints and this may have limited the amount of qualitative research that could be conducted, thus limiting the content validity of the study as well. Such studies may be useful for scientific purposes or as exercises involving training of students, but should be used in policy analysis and litigation only with the heaviest caveats. The second level of the hierarchy would involve studies that have achieved a fair amount of success in the rudimentary tests, but that either do not have the budget to support advanced testing or have not succeeded in passing advanced tests. Second-level studies may be usable in cost-benefit analyses, since normally the goal of such analyses is simply to determining whether the benefits of an intervention exceed the costs. Of course, unless benefits exceed costs by a fairly wide margin or vice versa, potential imprecision in second level studies may mean that the issue of whether benefits exceed costs remains open. Second level studies may be less applicable in litigation, where relatively precise estimates of value are needed to assess damages, but they may still be useful in preliminary damage assessments. Third level studies are studies that have conducted and achieved substantial success in passing advanced tests. Provided that such studies are judged to have a high degree of content validity as well, they would have the highest level of credibility for benefit-cost analysis and litigation.

The Clark Fork study report presents valuation equations. Independent variables included a scale item on the respondents' self-reported feelings about the importance of cleaning up the Clark Fork NPL sites; the sum of scale items on the importance of issues associated with groundwater, surface water, and terrestrial contamination in Montana; a measure of respondents' views about the likelihood of using the resources; a measure of respondents' rankings of the possible reasons for cleanup; and the self-reported degree to which respondents felt they should be responsible for paying for a cleanup. The regressions also included a series of variables reflecting respondent characteristics. These variables included income, age, gender, participation in recycling activities, proximity to the affected sites, and whether respondents had been members of or had contributed to an environmental organization in the year prior to receiving the survey.

The results of these regressions demonstrated that willingness to pay for cleanup was frequently related to these variables in ways that are consistent with economic theory. For example, the regressions consistently identified income as a significant positive predictor of willingness to pay. In addition, respondents living closer to the site expressed significantly higher values. Finally, the variables reflecting the importance of various aspects of the intervention were often significant. Individuals expressing higher levels of concern for cleaning up hazardous waste sites and/or higher levels of concern for cleaning up the Clark Fork NPL sites tended to express relatively high willingness-to-pay values. These rudimentary tests provide solid support for the construct validity of the study.

Moving to advanced tests, one can see from the survey results that many respondents prefer complete cleanup to partial cleanup of the Clark Fork sites. Thus, a scope test here would involve testing the hypothesis that complete cleanup had a higher reported value. The most rigorous form of such a scope test would involve a comparison of mean stated willingness to pay from two independent samples. In the Clark Fork study, the most rigorous scope test would involve a comparison between mean willingness to pay for complete cleanup from Version 1 of the survey instrument and mean willingness to pay for partial cleanup from Version 2 (as reported in Table 5-4). The Clark Fork study clearly does not pass this most rigorous scope test, obtaining nearly identical mean values for complete and partial cleanup before and after adjustments for protest zeroes, outliers, embedding, sampling issues, and feelings of responsibility.

This failure to pass the most rigorous scope test does count against the construct validity of the study, but should not be taken as damning evidence. The between-sample scope result could be a consequence of several factors. For instance, the test is predicated on the idea that there is, in reality, a difference in willingness to pay for the two levels of cleanup. Regardless of

the researchers' perception of differences between the two levels of environmental intervention, if nearly all respondents do not attach different values across the two levels of cleanup, a scope test could not be passed. In the same vein, if the difference between the willingness to pay for the two interventions is small, then detecting this difference might require very large samples, a potential problem that would be exacerbated in the Clark Fork study by the rather large gaps between values in the payment card.

The researchers do further analysis on the between-sample scope test using subsets of their samples. The embedding question was stated as follows:

Q30 Some people tell us it is difficult to think about paying to clean up just one site or even just one environmental problem. Would you say the dollar amount in Q28 you stated your household would be willing to pay is: (Circle number of best answer)

1 JUST FOR CLEANUP AT THE CLARK FORK RIVER BASIN. GO TO Q32.

2 PARTLY FOR CLEANUP AT THE CLARK FORK RIVER BASIN AND PARTLY TO CLEAN UP OTHER HAZARDOUS WASTE SITES.

3 BASICALLY A CONTRIBUTION FOR ALL ENVIRONMENTAL OR OTHER CAUSES.

4 OTHER (PLEASE SPECIFY)

Those who answered this question with "Other" were excluded from the data set and the between-sample scope hypothesis was again tested. This procedure is appealing because those who answered "Other" may be those who had the most difficulty with the CV questions to begin with. The subsample of remaining respondents shows a difference in values for complete and

partial cleanup. The difference is not very large, but is statistically significant at the 11 percent level in a one-tailed test (p. 5-41). Stated differently, the researchers can be 89 percent confident that the scope test is passed for this subsample of respondents. This supports the construct validity of the study, although not as much as passing the scope test with the full set of observations.

Further analysis relating to scope tests was conducted after the Clark Fork study final report was published in January 1995. The researchers focused on responses to Question 30 in combination with results from other questions in the survey. The new results are included Chapter 4 of the rebuttal report written by the state's experts in response to criticisms of their study by ARCO's economic experts. Four types of respondents were identified. These corresponded to the four different response categories in Question 30 were and, based on how they answered other questions in the survey, were labeled: (1) site specific, (2) site specific and generalized, (3) generalized, and (4) idiosyncratic. . Answers to other questions in the survey showed that Group 1 (i.e., those who chose the first response in Question 30) was more aware of the Clark Fork sites than the other groups, with Group 2 coming in second in this regard. Group 1 also generally felt that the other issues dealt with in survey Question 2 were less important than the other groups. In other ways as well, Group 1 seemed particularly strongly oriented toward the Clark Fork sites. Group 2 was also interested in the Clark Fork sites but had broader interests in other NPL sites and social problems than Group 1. Group 3 tended not to be very aware of the Clark Fork sites, and to care most about other problem. The rebuttal report, Chapter 4, p. 5, summarized Group 3's views this way, "Thus, while the upper Clark Fork sites are part of their general concerns, they do not particularly care about those sites and hence allocate a smaller portion of their WTP to those sites." Group 4 was the least aware of the Clark Fork sites and

was also the least concerned about the other problems they were asked to rate in Question 2. Half of them had zero adjusted values to begin with.

Values were highest for Group 1 and trailed off to relatively low levels for Group 4. This in itself is a positive result from the perspective of construct validity. Furthermore, Group 2 passed a between-subsample scope test at the .04 level in a one-tailed test. Group 1 had a substantial difference between the values of complete cleanup for those who filled out Version 1 and partial cleanup for those who filled out Version 2 ($9.30) but the difference was not statistically significant at conventional level, at least partly because of small sample sizes. Groups 3 and 4 failed the scope test.

The success with Group 2 and strong hints of scope sensitivity for Group 1 represent substantial evidence of construct validity. While not as good as passing the between sample scope test with the full samples (over half the total sample were in Groups 3 and 4 and thus failed to the between sample test), the Clark Fork study should be given credit for it progress toward passing advance construct validity tests. To simply write off the whole study as having failed the between-sample scope test and therefore consigning it to Level 2 in the construct validity hierarchy would be a much too extreme position given these partial successes.

Furthermore, another, albeit weaker, scope test involves a comparison of average values for complete cleanup to average values for partial cleanup for Versions 1 and 2 separately. This "within-sample" scope test is not as compelling as between-sample tests because, having read about both levels of cleanup, respondents may anticipate that the researchers are expecting different values commensurate with the different levels of cleanup. However, within-sample scope tests do provide information on construct validity despite this possible shortcoming. The Clark Fork study would pass this test.

Given the study's considerable success in passing rudimentary construct validity tests and its mixed, but positive, performance in passing scope tests, I would designate it a Level 3 study, but at the low end of Level 3. Clearly, a study that passes a between-sample scope test with flying colors would have to be considered stronger in construct validity. However, this shortcoming is partly remedied by the between-sample successes for subsamples. Furthermore, to ignore within-sample scope tests would not be justified, since they show that, given the two levels of cleanup in juxtaposition, many respondents recognized the differences and translated the differences into the values they expressed. That they were influenced by their perceptions of the researchers' expectations is possible, but there is no evidence on this topic one way or the other. It is somewhat encouraging that the residual values are so close between Versions 1 and 2. True, this could be coincidence. It could also be true that, as the report suggested (p. 5-11), "This [i.e., the similarity in residual willingness to pay across the two versions] may reflect that respondents have somewhat more measurement error in selecting a WTP amount for complete or partial cleanup than they have in determining the difference in value they assign when comparing two scenarios." Though weaker as evidence of construct validity than cross-sample scope tests, within-sample scope tests deserve to be treated as advanced tests in deciding whether to assign a study to Level 2 or Level 3.

CRITERION VALIDITY ASSESSMENT

Criterion validity studies, as I noted at the outset, involve field or laboratory experiments where values from actual transactions are used as standards for evaluating the accuracy of contingent values. It is tempting to draw sweeping conclusions from such studies. Indeed, the

rather presumptuous question "Does Contingent Valuation Work?" was part of the title of one of my earlier papers from a criterion validity study. Unfortunately, using such studies to judge whether CV "worked" in a given application like that of the Clark Fork study is not so simple. Perhaps the easiest way to demonstrate why is to examine a more recent actual criterion validity study in which the author was a participant, that by Champ et al. (1995).

Champ et al. (1995) conducted a criterion validity study involving removal of some old dirt roads from the North Rim of the Grand Canyon. These old roads allow unauthorized public access using motor vehicles into some remote areas there. Removal of the roads would reduce disturbance to wildlife and those attempting to enjoy wilderness recreation in these areas. Removal would also fulfill one of the requirements for designating the area as an official wilderness area. For these reasons, removal of the roads is a National Park Service goal. However, the Park Service lacks money to provide support for volunteers to carry out the work. Champ et al. asked a random sample of Wisconsin residents if they would actually donate money for road removal. Members of a second sample drawn from the same population were asked CV questions about their willingness to donate money. The actual donations then served as simulated market-like criteria for evaluating the validity of the CV donations. This study found a large potential upward bias in the CV responses. That is, people expressed willingness to donate more money for this purpose in the CV exercise than they would actually have donated.

Now, the issue is what I, as a reviewer, ought to infer from these results about how well CV performed in the Clark Fork case. In particular, is there a solid foundation in Champ et al. to conclude that the Clark Fork application of CV greatly overestimated the value that willingness of Montana residents to pay for clean up of the Clark Fork sites?

First, as was already noted earlier in this report, the temptation to reach generalizations from one or only a few studies about the overall validity of the CV method should be resisted. A study like Champ et al. has too many unique features, some of which are obvious and some of which are very subtle, that make it a special case, thus limiting its power to serve as a basis for sweeping generalizations. I believe that it would be very difficult to sustain an argument of the following form: (1) CV did not work well in Champ et al.; (2) therefore, CV does not work well in general; and (3) therefore CV did not work well when applied in the Clark Fork study. Champ et al (1995) had too many unique features to make such an argument viable.

The step from Conclusion (1) to Conclusion (2) would be especially questionable because Champ et al. (1995) used a donation vehicle. Donations are widely considered to be incentive incompatible valuation vehicles. Particularly when people actually have to pay, as they did in Champ et al.'s simulated market, donation vehicles invite respondents to "free ride," hoping that others will pick up the tab Preliminary inferences might be drawn about the efficacy of CV studies using donations vehicles but even this would be subject to additional research to replicate the result. Inferences to studies using other payment vehicles would be tenuous indeed.

If there was a closer match between the Grand Canyon road removal study and the Clark Fork study in terms of the procedures followed, a narrower criterion validity argument might be more tractable. The form of this argument would be that: (1) CV did not work well in the Grand Canyon Road removal study; (2) the Clark Fork study has much in common with the road removal study; (3) therefore, CV did not work well in the Clark Fork study. One does not have to look far to find many potentially relevant differences between the two studies that make such an argument doubtful. First and foremost, the Clark Fork study did not use a donation payment vehicle. Furthermore, Grand Canyon road removal is not very similar to a mining waste site clean

up in Montana. Champ et al. chose as subjects residents of a state far from where the intervention would occur. The Clark Fork study dealt with people who live much closer.

Potential confounding effects from donation vehicles limit the relevance of most of the criterion validity studies where nonuse values have been predominant, including not only Champ et al. (1995), but also Kealy et al. (1990); Duffield and Patterson (1992); and Seip and Strand (1992). Two potentially relevant studies did not use donation vehicles, Boyce et al. (1989) and Carson et al. (1986), but neither of these studies appears to provide many insights about the validity of the Clark Fork sites.[6]

More criterion validity studies are available if one is willing to consider use value studies, but here the overall analysis of Carson et al. (forthcoming), which I have already discussed in some detail, seems more directly useful than comparisons based on only a few studies. The conclusion there, it will be recalled, was that CV performed rather well on average.

Criterion validity studies are useful, but they take on the most relevance in the context of evaluating the overall validity of the CV method and then only when many studies are combined as was done in Carson et al. (Forthcoming). Disappointingly few insights about the validity of an individual application like the Clark Fork study are possible based on the criterion validity studies of nonuse values available at this time.

COMPARISON OF CLARK FORK STUDY WITH TWO OTHER STUDIES

[6]Boyce et al developed their criterion by offering to actually sell people house plants, where unpurchased plants would be killed. They found that CV overvalued such plants, but their CV exercise was not very well developed and used an open-ended format, rather than the payment card used in the Clark Fork Study. Carson et al. (1986) used CV to successfully predict the vote on a water quality referendum in California, thus supporting the validity of the method up to a point, but did not directly compare values.

The validity assessment of the Clark Fork study that was just presented was intended to stand on its own up to a point However, further understanding of my evaluation of the Clark Fork study can be added by applying the same validity assessment framework to two other studies and comparing them to the Clark Fork study. For these comparisons, I chose studies that were conducted in the context of the litigation over the Exxon Valdez oil spill and hence should have special relevance to natural resource damage assessment. One is the passive use (nonuse) value study done for the state of Alaska by Carson et al (1992) Many, including the NOAA Panel, have viewed that study to be of high quality. The other was performed by experts for Exxon and has been described in several places (Diamond et al. 1992; Hausman 1993; McFadden 1994). The research sponsored by Exxon did not address the damages from the Exxon Valdez oil spill directly, but rather estimated the values associated with avoiding logging in some wilderness areas in the West The purpose of study sponsored by Exxon was apparently to collect data that could be used to assess the overall validity of the CV method. The survey in that study involved a number of treatments that differed in terms of the format of the CV question and the number of wilderness areas to be logged. Depending on the treatment, logging would have occurred in from 1 to 57 wilderness areas in Colorado, Wyoming, Montana, and Idaho.

My assessment will focus on content and construct validity, since problems encountered when I tried to consider the criterion validity of the Clark Fork study are found here as well. WA will be used to symbolize the wilderness area study, CF to symbolize the Clark Fork study, and EV to symbolize the damage assessment for the Exxon Valdez spill. Each of the questions considered for CF will now be applied to WA and EV. Results are summarized in Table 2

Neither WA nor EV suffered from a great deal of theoretical ambiguity about the nature of the value (or values) they were trying to measure Both gave considerable explicit attention to

the theoretical definition of the true value they were seeking to estimate and received 4 out of 5 points on Question 1. By comparison, CF was rated one point lower than EV and WA in this regard because of my theoretical misgivings about its partitioning of resource values and because of its failure to consider the theoretical relationships between the true value of the damages and the concept of residual value that was used as a proxy, as discussed above.

More serious concerns arise about WA under Questions 2 and 3. It was apparently assumed that the only attributes relevant to potential respondents were the size of the areas to be logged, the percentage of each area to be logged each year, and the road building and use of heavy equipment that would be necessitated by logging. (Other wilderness areas are treated as substitutes in my assessment, not as attributes.) This is worrisome, since logging in wilderness areas would affect a host of other environmental attributes of potential relevance to study subjects. For example, nothing was specified about logging practices. Clear cutting is considered less desirable than selectively cutting by some members of the public, yet the issue was apparently not raised. Depending on logging practices, water quality in streams and downstream reservoirs could also be affected. Changes in water quality would in turn influence populations of fish and other aquatic organisms. In wilderness areas west of the Continental Divide, threatened and endangered species of salmon could be affected. Other wildlife populations would also be influenced and not all of the effects would be negative. Populations of large ungulates, for example, might benefit from logging. Potential effects of the logging on scenic vistas were not explained. CV studies conducted in the context of the spotted owl controversy showed that people are very concerned about the fate of old growth forests, yet the extent to which the proposed logging would involve old growth trees was not mentioned. There appears to be no recognition that opening up wilderness areas to logging might be viewed as a positive step by

some respondents since it would augment employment and regional incomes. So far as I can tell, the research involved no effort to document what these effects would be, to investigate which effects potential respondents would find relevant, to evaluate the current state of knowledge of potential respondents regarding the possible effects of logging, or to develop effective ways to communicate needed information to them. As nearly as I have been able to tell based on Diamond et al. (1992, p.22), the pretest consisted of administering by phone earlier versions of a mostly developed instrument with follow-up questions designed to help fine-tune the instrument. Such pretests are somewhat useful in improving communication but provide very limited information about more fundamental aspects of the scenario. Apparently, at least based on the written material from the study, no focus groups or verbal protocol exercises were held to document which of the affected attributes were important to potential subjects or to gain feedback regarding effectiveness of communication. Hence, WA is rated quite low under Questions 2 and 3. These low ratings reflect my doubts about whether respondents were well informed when they answered the CV questions.

CF and EV did much better in addressing these issues. Both were linked to large-scale efforts on the part of trustees to verify injuries to resources from the releases at issue and drew upon those efforts for information about the effects of the interventions being evaluated. Both involved extensive amounts of qualitative research to identify respondent-relevant attributes and learn how to communicate effects well. EV, in particular, worked intensively to address the issues under Questions 2 and 3 and received full points on these two questions. Its efforts in qualitative research were extensive. Its pretest and pilot survey instruments included questions regarding attributes and clarity of communications. Answers to these question led to the revisions of the final instrument designed to ensure that respondents were well informed when they came to

the valuation question. CF received fewer points on Question 2 than EV because its efforts to identify respondent-relevant attributes were adequate, but less elaborate than EV's efforts in that area. For Question 3 CF received fewer points than EV because of my concerns about the effectiveness of communications in that study and the scenario's lack of information about the effects of resource injuries as discussed above.

Turning to Question 4, like CF, neither WA nor EV explicitly reminded respondents about their budget constraints. However, all three were sufficiently stark in their emphasis on commitments of money to be satisfactory in this regard from my perspective. All three allowed respondents to reconsider their responses to the valuation question (CF through its embedding question), further emphasizing the need to consider carefully the possible payment of money to achieve the intervention CF asked respondents to consider alternative ways to pay for the intervention, which also served to emphasize that commitments expressed in the CV question would, if actually paid, deplete the respondent's budget. EV went further in its treatment of this issue than either CF and WA by mentioning that some respondents would vote against the referendum because they could not afford the cost.

CF and EV are roughly similar in the amount of information they present about environmental substitutes (the other issue touched on under Question 4). For example, both presented maps that clearly showed that large percentages of the areas of the respective states were undamaged. Both included questions early on that dealt with spending public money on activities other than the environmental projects that were ultimately valued (i.e., spending on potential substitute public activities). The WA instrument did stress the total number and acreage of wilderness areas in the four-state area and the number to be logged. Beyond that, however, WA was very skimpy on details about undamaged substitutes. Except for saying that at least one

would be located in the respondent's state, neither the location of other wilderness areas (7, 8, or 9 depending on the treatment) to be logged nor the total numbers of acres involved was mentioned.[7] Hence, respondents did not know which wilderness areas would remain undamaged. Furthermore, no maps or other visual aids were provided to help respondents visualize the extent of undamaged substitutes.

Combining my evaluations of how these studies treated budget constraints and substitutes, I assigned WA 3 points, while both CF and EV were assigned 5 points.

Turning to Question 5, EV had a carefully designed and very detailed context for valuation. Furthermore, the referendum format used by EV is widely considered to have theoretically satisfactory incentive properties. The incentive properties of the CV questions in WA and CF were less clear. Both EV and WA used a tax vehicle. This may have introduced a downward bias in valuation responses because of the unpopularity of taxes generally. Letting respondents choose their vehicles, as was done in CF, appears to be more neutral in this regard. Considering all these aspects, I assigned WA 6 points and EV, 8 points for context. This compares to CF's 7 points on this item.

The acceptability and believability of the scenarios come under Question 6. Recall the difference. Quoting Appendix B, "A study subject accepts the scenario when he or she implicitly agrees to proceed with the valuation exercise based on the information and context provided." A scenario is "believed" to the degree that respondents expect their responses to the CV question to actually affect how much of the environmental amenities in question they will receive and how much they will actually pay if the intervention is adopted. A scenario is acceptable, yet not

[7] The treatment where all 57 wilderness areas were to be logged was an exception.

believed when respondents agree to engage in "what if" exercises where they know that the scenario is completely hypothetical.

One of the goals of the extensive efforts that went into the design of the EV survey instrument was to develop a final scenario that would be both acceptable and believable to respondents. Debriefing questions verified that this goal was largely achieved. EV was assigned 8 points rather than the full score of 10 points on Question 6 because the tax vehicle may have impeded acceptance of, and belief in, the scenario. That a federal income tax would be levied for one year to establish an escort ship program for oil tankers in Prince William Sound may have seemed a bit implausible to some respondents given the lack of precedent for such a tax. Also, the valuation questions explicitly stated that the intervention would cost the household in question a specified amount in taxes. This was bound to raise questions in the minds of some respondents about how the interviewer could know the exact implications of the intervention for that household's taxes.

The WA ran the risk of severe difficulties with regard to acceptability and believability. Logging of even one wilderness area would be a major environmental battle. A proposal to log all 57 wilderness areas in the four states would have generated a tremendous furor that would have occupied the national and regional press endlessly. McFadden (1994, p. 696) pointed out, "This resource issue was chosen because at the time of the study in 1990 there was active discussion in Congress and in the media in the western U.S. regarding logging on government lands, making logging familiar to many respondents." However, logging on federal lands is one thing, logging in designated wilderness areas is quite another. There are two aspects of WA that are likely to seem especially implausible to many respondents, that a specified number of wilderness areas (7, 8, or 9 depending on the survey version) had already been slated for logging

without their realizing it and that all 57 wilderness areas in four states were being considered for logging. Not much is known about how members of the public are likely to respond when confronted over the phone with a major policy change that they find implausible and about which they have not previously heard or read. Designers of the survey apparently were aware that respondents might have a negative reaction to the logging proposal. Interviewers were instructed to tell respondents expressing concern about the issue that logging in wilderness areas ". . . is not currently being considered by lawmakers." Outrage, suspicion about the motives for the survey and its true goals, refusal to participate in the survey or, if they proceed, to take the survey seriously, and curiosity about details (in this case not provided) are all plausible reactions. Such reactions create doubts about the quality of resulting data. Diamond et al. (1992 p. 22) reported, "Overall, the response rate was 62%." Unfortunately, they do not elaborate on how this was calculated or how the non-response was broken down. If this number expresses interview completions as a percentage of verified residential telephone contacts, it seems very low and could reflect scenario rejection Like EV, WA's use of a tax vehicle could have reduced both acceptability and believability.

By comparison, CF comes closer to EV in terms of acceptability and believability. Many members of the sample had heard of the Clark Fork sites and probably were not surprised by being surveyed on the topic nor would they have considered the whole subject of cleanup a substantial departure from prior beliefs about existing policy. Allowing respondents to choose their payment vehicles may have increased the acceptability of the scenario, while making it less believable.

For scores on Question 6, WA was assigned 4 out of the possible 10 points, while EV received 9 points. These scores compare to 8 points assigned to CF.

Moving to Question 7, comparing survey questions other than those directly related to valuation across the three studies is difficult. All three asked a variety of other questions, which in some respects were quite similar. In other respects, however, there were major differences. This caveat should be borne in mind in considering my ratings for this question. I assigned EV 10 point, CF 8 points, and WA 6 points. EV received the 10 points because it gathered a great deal more data. Particularly noteworthy in the EV survey were several questions exploring respondent reactions to the scenario and detailed interviewer debriefing questions. CF included several questions potentially useful in construct validity assessment for which there were no counterparts in WA. These included questions about how important respondents felt it was to clean up the various sites, how satisfied they would feel with various levels of cleanup, how responsible they felt for cleaning up hazardous waste sites, counties of residence (used in calculating distance from the Clark Fork NPL site), importance to respondents of toxic cleanups compared to other issues, and possible future use of Clark Fork resources. Thus, CF is rated higher than WA on this item.

Regarding survey mode (Question 8), many practitioners of CV agree with the NOAA panel in preferring personal interviews. EV involved personal interviews conducted by a leading national survey firm. This greatly enhances its content validity as I recognized by awarding it 10 points. Telephone interviews like those employed by WA are considered inferior for nonuse value studies. Not only is the amount of verbal material that can be presented effectively in a telephone interview limited, but it cannot be supported by visual aids such as maps, diagrams, and photographs. Also, constraints on time to complete the survey may be more severe for telephone surveys than for either mail or in-person surveys. Even if the researchers conducting WA had done more to document which attributes of wilderness areas were important to potential respondents and had thoroughly documented the effects of logging on those attributes, I believe

the choice of a telephone survey mode would have prevented them from adequately communicating what they had learned. My pessimism about the potential for telephone CV surveys led me to assign WA only 2 points for survey mode.

Regarding Question 9, WA involved three formal pretests with fairly large samples and included debriefing questions to attempt to investigate whether respondents understood the scenario. I assigned it 1 point less than CF for this item because CF included not only pretests but also verbal protocols. EV received a full 5 points because of the potentially more effective, very extensive procedures it followed, which involved application of qualitative research tools as well as pretests and pilot studies.

Turning to Question 10, WA and EV both followed sound survey execution procedures and received high scores. EV received the full 10 points. Sampling was accomplished using a multistage area probability sample. This sample was constructed by first randomly choosing 61 counties or groups, then randomly selecting 330 blocks within counties, and finally selecting 1,600 residences from the selected blocks. This type of sampling provides perhaps the best coverage of the population. Standard, high quality procedures were followed in gathering data. EV reported a response rate of 75.2 percent. This was calculated as completions divided by the number of potential interviews less non-English speaking households and vacant residences. The WA sample was obtained using random digit dialing. The coverage of this type of sample is quite good, the major deficiency being the loss of households without telephones. WA reported that up to 10 follow-up phone calls were made to attempt to gain a response, which seems quite laudable. Depending upon how it was calculated, the 62 percent reported a response rate in WA may raise questions about the ultimate effectiveness of these procedures. WA was assigned 7 points, reflecting its low response rate. For reasons already noted, CF received 8 points on this item.

Regarding econometric procedures (Question 11), the analysis of McFadden (1994) using the WA data was outstanding. The design of the analysis was driven by economic theory, was very thorough, and was commendable in exploring a number of functional forms as well as a non-parametric procedure. Thus, I assigned WA 10 points for econometrics. CF and EV were adequate in this regard as well but less innovative and were assigned ratings of 8.

Turning to Question 12, the written materials from EV were very thorough and explicit regarding the procedures followed and the results of the analysis, warranting full points. WA did not have a complete report for easy reference, but details could be pieced together fairly well from the various papers that have been written. I assigned it a 4 for reporting, the same as CF received on this item.

Total points on the 12 detailed question are combined in Table 2. The scores summed to 53 points for WA and 96 points for EV, compared to 75 points for CF.

Turning to Question 14, no other issues arose in my reviews of WA and EV.

Question 15 asks for an overall quality rating. It will be remembered that I assigned CF a rating of Good I see every reason to rate EV as Excellent. Though I was quite critical of WA on some points, it also had some good points and I rated it as Fair

Table 2: Comparative Content Validity Ratings for the Wilderness Area (WA), Clark Fork (CF), and Exxon Valdez Oil Spill (EV) studies

Question	WA	CF	EV
1 True value defined?	4	3	4
2. Attributes identified?	2	8	10
3. Effects documented and communicated?	2	6	10
4. Aware of budget constraints and environmental and other substitutes?	3	5	5
5. Context specified and incentive compatible?	6	7	8
6. Scenario accepted? Scenario believed?	4	8	9
7. Other questions?	6	8	10
8. Survey mode appropriate?	2	6	10
9. Potential flaws identified?	3	4	5
10. Survey procedures?	8	8	10
11 Econometrics	10	8	10
12. Written materials adequate?	4	4	5
Total Points	54	75	96

What about construct validity? I would rate CF and EV as roughly equivalent in terms of their overall construct validity. That is, both are toward the lower end of Level 3. Unfortunately, EV was also done before the advanced construct validity tests began to be taken so seriously. Hence, it has no between-sample scope test. While it did not fail such a test, neither did it pass one. EV does have a very strong valuation equation. An indirect, within-sample scope test of sorts can be conducted by examining this equation (Carson et al. 1992, p. 5-108). One set of

dummy variables represents respondents' views about the potential seriousness of oil spills if the escort ship program, which served as the intervention, was not adopted. These variables can be interpreted as respondents' perceptions of the scale of the injury without the program. Those who felt that the injuries would be a great deal more or somewhat more than those caused by the Exxon Valdez spill were willing to pay significantly more for the program, while those who felt that the injuries would be less or that there would be no injuries were willing to pay less. Likewise, those who felt that the escort ship program would not be very effective in preventing injuries or would not be at all effective were willing to pay less on average than those who thought the program would work. Not only are all of these relationships significant at the 10 percent level, but they have the correct relative magnitudes. This is a somewhat stronger within-sample scope test than that in CF because it does not involve direct comparisons of scenarios but respondents' perceptions regarding a single scenario. Thus, it avoids the fear that multiple scenarios will lead respondents to give answers in accordance with expectations of the researchers.

WA, on the other hand, appears to have much lower construct validity. Valuation equations presented in McFadden (1994) show mixed results. Potential explanatory variables are sometimes significant predictors of willingness to pay, but at other times are not. McFadden tested the hypothesis that income had the same effects across several treatments and could not reject it, a positive result. However, scope tests were soundly failed as were other tests based on theory (McFadden 1994; Diamond et al. 1992). I would assign WA to the lower end of Level 2, implying that its results might be useful for scientific purposes, but would need to be applied only with heavy caveats, if at all, in policy analysis and damage assessments.

COMMENTS ON THE SCIENTIFIC STATUS OF CONTINGENT VALUATION

The courts are often asked to evaluate the merits of expert scientific testimony. Paraphrased to simplify a bit, the following four guidelines have been suggested. (113 5.Ct.2786 (1993))

1. Does the method generate hypotheses and can these hypotheses be tested?

2. Has the method been subject to peer review and publication?

3. Is the rate of error known and are there standards for controlling the technique's operation?

4. Has the method obtained a general acceptance among a relevant scientific community?

Of course, I am in no position to express opinions on legal issues. Nevertheless, I thought it might be helpful to give one economist's answer to these questions as they apply to CV.

On the first question, there seems to be no doubt. The goal of construct validity assessment is to test hypotheses about relationships between results from CV studies and expectations from economic theory. Theory-driven hypothesis testing is central to CV.

Regarding peer review and publication, a number of CV studies have appeared in prominent, peer-reviewed economic journals including the American Economic Review, the Quarterly Journal of Economics, the American Journal of Agricultural Economics, Land Economics, and the Journal of Environmental Economics and Management.

Turning to the third question, errors can exist on two levels. Statistical error (the error associated with sampling) can certainly be evaluated for contingent values and confidence intervals around estimates of mean values can be derived from the same statistics, as can estimates of the standard deviation of contingent values for populations of individuals. In this sense, error

levels for contingent values are as quantifiable as those for any other economic estimates. On a more difficult level, the direction and magnitude of the bias (defined as the observed value minus the true value) in any given value estimate based on CV is not known since true values are not observable. However, this is true for all economic value estimates. As to whether the rate of error can be controlled, errors can be minimized by designing and executing studies with high content validity, a point of view that I share with the NOAA Panel. Furthermore, construct validity tests provide important clues about which studies are more likely to be accurate and which are more likely to contain large errors. Criterion validity studies should eventually provide further insights about the error levels in nonuse valuation studies, but as noted earlier, the evidence so far is not very helpful.

Regarding whether CV is generally accepted, the history of the CV method is instructive. Prior to the Exxon Valdez oil spill, the method was not the matter of intense debate that it is today. Summer meetings of the American Agricultural Economics Association and winter meetings of the Allied Social Science Association devoted sessions to recent papers on the topic each year These sessions were almost invariably well attended. A large group of researchers from Agricultural Experiment Stations across the country formed Western Regional Committee W-133, which held meetings annually focused on nonmarket valuation, attracting not only its members but dozens of researchers and government economists. Many of the papers presented there focused on the CV method. The result of all this effort was increasing acceptance of the method by academic resource economists and federal agencies, who saw the technique as a promising new approach to measuring previously unquantifiable benefits and costs. The U.S. Environmental Protection Agency; Department of the Interior agencies including the Bureau of Reclamation, the Fish and Wildlife Service, and the Bureau of Land Management; the U.S. Army

Corps of Engineers; and U.S. Department of the Agriculture agencies such as the Forest Service and the Soil Conservation Service all authorized use of CV for policy analysis. Scholarly interest has spread beyond the U.S. Studies have now been completed in Canada, Australia, New Zealand, the Republic of China, Great Britain, Germany, Norway, Italy and other countries.

Following the Exxon Valdez spill, and largely because of the proposed use of the method in natural resource damage assessment cases, prominent mainstream economists took an active role in critiquing the method and attacking its validity. This has had the benefit of focussing the CV debate on validity issues. Some have staked out extreme positions on both sides of the issue, but most, including the economists on the NOAA Panel, have focused attention on the determinants of where the method will work well and where it will not. This new focus seems to be moving in the direction of seeking to identify specific actions within the control of the researcher that will enhance the validity of the results and on the interpretation of evidence from theory-driven hypothesis testing Interest in research on the topic is growing rapidly, as is evidenced by a new program of support sponsored jointly by the Environmental Protection Agency and the National Science Foundation to explore various issues in environmental valuation, including CV. This in itself is significant evidence that many at high levels of academia and government continue to consider the method useful and promising. The validity of the CV remains the focus of heated debate. However, among those most knowledgeable about the technique, there is substantial support.

SUMMARY AND CONCLUSIONS

This report began by considering the overall validity of the CV method. Though everyone agrees that more research is needed, many of those who have stepped back and tried to gain an overview of the validity of the method have concluded that the method has demonstrated enough validity to make well-done studies useful in policy analysis and damage assessment. This was the conclusion of the NOAA Panel on Contingent Valuation and the bulk of the economists who have worked most directly on the development of the method over the past 30 years. It is shared by several federal agencies including the U. S. Environmental Protection Agency, various agencies within the USDA and the Department of the Interior, and the U.S. Army Corps of Engineers. Valuation equations from many studies have demonstrated that CV data agree with prior expectations based on economic theory. Furthermore, many studies have passed scope tests. Given the validity that most economists assign to values from applications of revealed preference methods, the comparisons reported by Carson et al. (Forthcoming) are particularly potent evidence. They have demonstrated that CV shows a rather strong tendency to produce values comparable to revealed-preference values in cases where the two major approaches have been applied side-by-side. All this tends to support the hypothesis that responses to CV questions are rooted in real world processes akin to those modeled in economic theory and hence that contingent values deserve to be called economic values.

It is true that the debate over the validity of the method has heated up of late. Economists and other scholars with sterling credentials have expressed grave doubts about the CV method. The context within which these doubts have been expressed is important to understanding them. With very few exceptions, the critics were not much involved with CV until they became experts for Exxon. I do not accuse these people of scientific dishonesty. But, I do believe that the inherent conservatism regarding new methods of those at the center of a well-established

discipline like economics, in combination with the pressures of high visibility litigation involving hundreds of millions, if not billions, of dollars, has affected how they view the matter. The critics of CV are good economists and they raise interesting issues. Their criticisms of CV deserve answers but those answers can only come through a long and arduous process of research. In the meantime, and particularly given the context of their critique, their strong propensity to throw out the proverbial "baby with the bath water" should be resisted.

That many studies have realized a measure of success does not mean that an individual study has necessarily achieved a satisfactory level of validity. Each individual study must still demonstrate high levels of content and construct validity and the Clark Fork study is no exception. The Clark Fork study seemed to me to have been quite solid in most of dimensions of its design and execution. It avoided serious problems in identifying respondent-relevant attributes, calling attention to budget constraints and substitutes, developing a context for valuation, devising a scenario respondents would find acceptable, asking sound questions other than the CV question itself, seeking out flaws in the survey through qualitative research and pretests, executing a sound survey, analyzing the data, and reporting results.

I did question some aspects of study procedures. It seemed to me that the survey required respondents to read and digest a great deal of material in order to deal effectively with the CV question. This concern was obviated somewhat by the $20 incentive, which may have helped to encourage them to work hard at it. Still, the problems that some respondents may have had in dealing with the lengthy scenario led me to assign only 6 out of 10 points each for communication and choice of survey mode. Six out of 10 points is not a bad score, but reflects my judgement that the scenario was not as crisp and simple as it should have been for a top quality mail survey. The study was also docked slightly for neglect of theoretical issues associated with apportionment

of values and failure to recognize and address the inconsistency between the complete cleanup scenario and baseline conditions (i.e., conditions with no releases of toxic chemicals). A scientific study should be tight theoretically and this study left some loose ends. While this was "bad form," so to speak, I doubt that neglecting these theoretical matters did any great amount of damage to the accuracy of the results from the study. I also made a slight deduction for problems with the context relating to the timing of cleanup. Had it not had these shortcoming, my overall content validity rating for this study would have been somewhere between "very good" and "excellent." Given these concerns, I still would rated the content validity of this study as "good."

Since content validity assessment identified some possible difficulties particularly in the area of communications, I entered the construct validity phase of the assessment with some concerns about how valid the final results would be. Had the validity testing failed to show positive results, the conclusion would have been that the Clark Fork study lacked validity, perhaps because of ineffective communications. However, the outcome of the construct validity assessment was quite reassuring. Valuation equations and success in the some between-sample scope tests with partial samples as well as the within-sample scope tests indicated that respondents dealt sufficiently well with the information that was communicated to them to answer the survey in ways that made sense from the perspective of economic theory. This supports interpreting the Clark Fork study results as valid estimates of the true values.

Above this rather positive conclusion is one potential important remaining concern, the failure of the Clark Fork study to pass the between-sample scope test for the full sample. While I would feel a little more confident about the Clark Fork study if it had passed such a test, I also believe that it would be a mistake to allow this one shortcoming to overshadow its many positive

results. Furthermore, when comparing subsample values, the Clark Fork study shows solid evidence of sensitivity to scope.

Comparing the Clark Fork study to two other well-known studies was designed to place it in a broader perspective. The Clark Fork study compared well to the leading-edge (and very expensive) Exxon Valdez study nonuse value study of Carson et al. and, in my opinion, was clearly superior to the wilderness area study. This supports the conclusion that, though it may not be quite as strong as the study of the Exxon Valdez spill, the Clark Fork study should command a relatively high level of confidence.

In my opinion, then, results from the Clark Fork study have sufficient validity to be used as measures of the true values of Montana residents for partial and complete cleanup at the Clark Fork sites.

REFERENCES

Boyce, R.R., T.C. Brown, G.D. McClelland, G.L. Peterson, and W.D. Schulze. 1989. "Experimental Evidence of Existence Value in Payment and Compensation Contexts." Paper presented at the joint meetings of the Western Committee on the Benefits and Costs of Natural Resource Planning (W133) and the Western Regional Science Association, San Diego, CA.

Cambridge Economics, 1992. "Contingent Valuation: Tests of Validity and Reliability." Papers from a symposium held at Cambridge Economics, Inc., Washington, D.C., April 2 and 3.

Richard T. Carson, 1995. "Contingent Valuation Studies and Insensitivity to Scope." Revised Version of Paper Was Presented at the International Conference on Determining the Value of Non-marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods, Bad Homburg, Germany, July 1994.

Carson, R.T., W. Michael Hanemann, and Robert C. Mitchell. 1986. "Determining the Demand for Public Goods by simulating Referendums at Different Tax Prices." Manuscript, University of California, San Diego.

Carson, R T., R C. Mitchell, W.M. Hanemann, R.J. Kopp, S. Presser, and P.A. Ruud, 1992. "A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill." Prepared for the Attorney General of the State of Alaska.

Carson, R.T., and N.E. Flores, 1992. "A Second Look at 'Does Contingent Valuation Measure Preferences?: Experimental Evidence?'" Unpublished paper Department of Economics, University of California, San Diego.

Carson, R T., N E. Flores, K.M Martin, and J.L. Wright, forthcoming. "Contingent Valuation and Revealed Preferences Methodologies: Comparing the Estimates for Quasi-Public Goods." *Land Economics*.

Champ, Patricia A., Richard C. Bishop, Thomas C. Brown, and Daniel W. McCollum. 1995. "A Comparison of Contingent Values and Actual Willingness to Pay Using a Donation Provision Mechanism with Implications for Calibration." Unpublished paper, Rocky Mountain Forest and Range Experiment Station, USDA Forest Service, 3825 E. Mulberry Street, Fort Collins, CO 80524

Desvousges, William H., F. Reed Johnson, Richard W. Dunford, Kevin J. Boyle, Sara P. Hudson, and K. Nicole Wilson *Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy*. Research Triangle Institute Monograph 92-1. Research Triangle Park, NC: Research Triangle Institute.

Diamond, Peter A., and Jerry A. Hausman. 1994. "Contingent Valuation: Is Some Number Better Than No Number?" *Journal of Economic Perspectives* 8(4):45-64.

Diamond, P., J. Hausman, G. Leonard, and M. Denning, 1992. "Does Contingent Valuation Measure Preferences? Experimental Evidence." Paper presented at the Cambridge Economics, Inc. symposium titled "Contingent Valuation: A Critical Assessment," Washington, D.C., April 2 and 3.

Duffield, J.W., and D.A. Patterson. 1992. "Field Testing Existence Values: Comparison of Hypothetical and Cash Transaction Values." Paper presented at the joint meetings of the Western Committee on the Benefits and Costs of Natural Resource Planning (W133) and the Western Regional Science Association, South Lake Tahoe, Nevada.

Fischoff, B. and L. Furby. 1988. "Measuring Values: A Conceptual Framework for Interpreting Transactions with Special Reference to Contingent Valuation of Visibility." *Journal of Risk and Uncertainty* 1:147-184.

Hanemann, W. M. "Valuing the Environment through Contingent Valuation." *Journal of Economic Perspectives* 8(4):19-43 (Fall).

Hausman, J., ed., 1993. *Contingent Valuation: A Critical Assessment*, Amsterdam: Elsevier.

Hausman, J., 1995. *Expert Report for the Atlantic Richfield Company in State of Montana v. Atlantic Richfield Company (No. CV-83-317-HLN-PGH).* July 13, 1995.

Hoehn, J.P., and A. Randall, 1987. "A Satisfactory Benefit Cost Indicator from Contingent Valuation." *Journal of Environmental Economics and Management"* 14(3):226-247.

Kahneman, D., and J. Knetsch, 1992. "Valuing Public Goods: The Purchase of Moral Satisfaction," *Journal of Environmental Economics and Management* 22:57-70.

Kealy, M J., M. Montgomery, J F. Dovidio. 1990. "Reliability and Predictive Validity of Contingent Values: Does the Nature of the Good Matter?" *Journal of Environmental Economics and Management* 19:244-263.

McFadden, D., 1994 "Contingent Valuation and Social Choice," *American Journal of Agricultural Economics*, 76:689-708.

Mitchell, R C., and R.T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method.* Washington, D.C.: Resources for the Future.

Paris, Q., M.R Caputo, and G.J. Holloway, 1993. "Keeping the Dream of Rigorous Hypothesis Testing Alive," *American Journal of Agricultural Economics*, 75:25-40.

Samuelson, Paul. 1954. "The Pure Theory of Public Expenditure." *Review of Economics and Statistics* 36:387-389.

Seip, K. and J. Strand. 1992. "Willingness to Pay for Environmental Goods in Norway: A Contingent Valuation Study with Real Payment." *Environmental and Resource Economics* 2(1):91-106.

APPENDIX A:

MEASURING NON-USE VALUES: THEORY AND EMPIRICAL APPLICATIONS

MEASURING NON-USE VALUES: THEORY AND EMPIRICAL APPLICATIONS

by

Richard C. Bishop, Patricia A. Champ, Thomas C. Brown and Daniel W. McCollum[1]

In keeping with our assignment, this paper will focus on measurement. We will devote far less time to the microeconomic theory of non-use values than many readers might expect. Non-use values are now well entrenched in the theory of the consumer (Krutilla 1967; Randall and Stoll 1983; Madariaga and McConnell 1987; Boyle and Bishop 1987; Smith 1987; Bishop and Welsh 1992; Freeman 1993). We shall want to briefly review the theory as a foundation for what will follow, but need not dwell on it at length. The central focus will instead be on measurement.

To date, contingent valuation (CV) is the only tool for estimating non-use values that has a substantial following among researchers. As most readers are no doubt aware, CV is currently the subject of debate in the U.S. Respected researchers currently disagree about whether CV can produce sufficiently accurate values to support damage assessments and policy analyses. This debate may have already begun in other countries and will almost certainly intensify there as applications of CV expand. The premise of this paper is that progress in this debate is impeded by the lack, within economics, of a well-articulated <u>theory of measurement</u> that is consistently applied across valuation studies. Mitchell and Carson (1989), drawing on the psychological theory of testing (Bohrnstedt 1983), have laid a foundation for such a theory. This paper will attempt to elaborate on their work with specific reference to non-use values.

In psychology, such abstract concepts as intelligence have long been the subject of research. Tests, such as IQ tests, are a major tool in psychometrics. We economists today face problems that are in many ways comparable to the problems that pioneers in psychology must have faced. We have the abstract concepts of willingness to pay (WTP) and willingness to accept (WTA) that we seek to measure. WTP and WTA, like intelligence, are

fundamentally abstract, theoretical concepts that are not directly and fully observable in the real world. Both economists and psychologists must try to infer something about the magnitudes of their constructs from data that are thought to be indicative of those magnitudes. In this endeavor, both must rely--implicitly or explicitly--on a theory of measurement to guide research design and the interpretation of results. In fact, one might reinterpret CV questions as "WTP tests" analogous to IQ and other psychological tests. Thus, an economic theory of measurement has much to learn from the psychometric theory of measurement.

We begin by defining theoretical WTP, a concept that will play a role here comparable to that of a "true value" in psychometrics. That is, the validity of observed WTP can only be evaluated with reference to inherently unobservable theoretical WTP. Once theoretical WTP has been defined, we will consider alternative forms of evidence for estimating it. Values estimating using market data have long been considered credible, but this does not alter the fact that they too are estimates of unobservable theoretical WTP and thus share with CV the problems associated with validity assessment. To address the thorny relationships between observable and unobservable values, we will consider the dual criteria of reliability and validity and adapt these concepts to reflect the goals of CV studies. We then turn to a triad of concepts that might be termed the "Three C's": content validity, construct validity, and criterion validity. These concepts represent mutually reinforcing strategies for assessing validity. Much of the paper focuses on how these strategies can be applied to assess the validity of individual CV studies. We will also emphasize that validity at the level of the individual study is a prerequisite for trying to draw empirical conclusions about the validity of the CV method as a whole.

Theoretical WTP

Let us work with WTP, always bearing in mind that WTA is an alternative definition of economic value with equal theoretical standing. To define

theoretical WTP, let us consider the effects of an "intervention" in the economy on the welfare of a theoretical consumer. Such interventions may take the form of proposed governmental policies, projects, or regulations that will in some way affect the natural environment.[2] Or, interventions might be releases of oil or a toxic substance with adverse environmental repercussions. In addition to environmental services, interventions may change other economic parameters such as the prices paid or the income received by consumers. Some environmental parameters may be relevant because the consumer "uses" the environment services in question. For example, the intervention may affect fish populations where the consumer goes fishing. Other parameters may be relevant to the consumer for reasons other than personal use. For example, the consumer may gain utility from being altruistic toward others who are users or toward animals or toward future generations. It is through effects on environmental and other economic parameters that interventions generate environmental "values," both use values and non-use values.

Suppose that the consumer maximizes a "well-behaved" utility function subject to a budget constraint of the standard form. To keep it simple, assume away time, uncertainty, and other complicating factors. Let us represent the consumer's indirect utility function by

$$v(P, d, Y)$$

where P symbolizes a vector of market prices, d symbolizes the status of an environmental resource that would be affected by the intervention, and Y symbolizes income. Let P_o, d_o, and Y_o represent the levels of these parameters in the absence of the intervention and P_w, d_w, and Y_w represent the levels of the parameters if the intervention is completed. Let WTP_t represent the theoretical WTP associated with the intervention. If the intervention has a positive effect on the consumer, that is,

$$v(P_w, d_w, Y_w) > v(P_o, d_o, Y_o),$$

then theoretical willingness to pay is defined implicitly by

$$v(P_w, d_w, Y_w - WTP_t) = v(P_o, d_o, Y_o).$$

On the other hand, if the intervention will make the consumer worse off, that is,

$$v(P_w, d_w, Y_w) < v(P_o, d_o, Y_o),$$

then the implicit definition becomes

$$v(P_w, d_w, Y_w) = v(P_o, d_o, Y_o - WTP_t).$$

Where the intervention leads to an improvement in welfare, WTP_t is most naturally interpreted as Hicksian compensating surplus, while it should be interpreted as Hicksian equivalent surplus if welfare will decline.

WTP_t as just defined must be viewed as a theoretical construct, a useful scientific fiction describing what would be measured under idealized circumstances. We economists use the WTP_t construct with such ease and confidence that it takes on a reality of its own in our minds, though it does not exist in reality. The utility maximization framework is a highly stylized version of how real human beings think and behave.

Consider, for example, how we conceptualize the process by which utility is maximized. Do we really believe that human beings decide at one point in time what they intend to consume during the next period of time? What unit of time are we thinking of, a month, a year, a lifetime? Is the "utility counter" set at zero on January 1 of each year? Human choice must, in reality, be a much more dynamic, complex process than the simple theory allows.

The very concept of "consumer" involves an abstraction from reality. Do we mean that consumption decisions are made by individual human beings acting in isolation from those around them? This does not seem square with reality.

Groups of individuals (couples, families, unrelated individuals who share dwellings) make joint decisions, often subject to constraints that involve multiple incomes. Thus, it is not unusual to find discussions that refer to the consuming unit as a "household." However, more rigorous treatments of the subject refer only to individuals as the basic consuming unit because, if group decision processes were permitted, basic assumptions about preferences such as completeness, transitivity, and continuity would be much more difficult to justify. These assumptions underlie the utility functions needed to define WTP_t (and theoretical WTA) in the usual manner and to develop hypotheses about the relationships between values and other economic parameters. To the extent that the theoretical concept of "the consumer" has no full counterpart in the real world, neither does WTP_t. WTP_t is in principle unobservable because it does not exist in reality.

Why deal in such abstractions at all? People appear to be willing to pay some amounts, but not other amounts, for conventional goods and services. It is not much of a leap to infer that real people might also be willing to pay something to obtain environmental amenities and avoid environmental "bads." Why not stop there, defining theoretical WTP as the maximum willingness to pay of real people? The answer is that such an approach would not be very rich theoretically speaking. It would not readily yield testable hypotheses about how willingness to pay, thus defined, might be related to the other economic parameters confronted by these same real world people. Whether one wishes to consider relatively simple relationships (e.g., the relationship between WTP and income) or more complex ones (e.g., the relationships between WTP for two or more commodities or WTP for public goods) abstract "modeling" of economic behavior is necessary. Abstraction makes it possible to explore seemingly fundamental relationships without having to deal with all the complexities and ambiguities inherent in the behavior of real world people. Consumer theory and welfare economics are useful in designing empirical studies and interpreting results. Gains of this kind are only realized, however, by assuming away potentially important parts of reality.

Thus, WTP_t gives rich conceptual meaning to the otherwise rather loosely defined concept of economic value. To implement the concept requires real world evidence. Having defined the theoretical construct that is to be the objective of measurement, the next step is to consider possible sources of data that could be used to estimate WTP_t.

Alternative Sources of Evidence Regarding WTP

Market data have been central to efforts to quantify economic values. This is the so-called revealed preference approach to value estimation. People spend money on goods and services, and such spending has great credibility among economists as evidence of values. Without quarreling at all with the credence that economists place on market data, it is worth emphasizing that market values are not direct observations on WTP_t. This is the case for both conceptual and empirical reasons.

From a conceptual perspective, to admit that economic theory is abstract is to admit that it may not and probably will not fully represent decision processes of real world consumers in the marketplace. People making market choices will hopefully engage in processes and apply criteria like those depicted in theory. Nevertheless, to say that theory is abstract is to say that what we shall term "other factors," factors not considered in the theory, may also impinge on consumer choices. Other factors could include, for example, the influences of group decision making within the household and strategies devised by real human beings to cope with uncertainty. To the extent that other factors enter into market choices, market values may diverge from WTP_t.[3]

In addition to theoretical concerns about interpreting market values as WTP_t, it must also be remembered that empirical estimation of demand and supply relationships inevitably involves simplifying assumptions. Errors in variables, model specification errors, missing data, extrapolations beyond the range of the data, and other potential sources of error may prove unavoidable. Benefit-cost analysis almost invariably involves extrapolations into the future

of values inferred from past behavior, a process that is far from perfect. Thus, even if real world consumers did behave exactly like the consumer as depicted in economic theory, econometric errors would make estimated market values imperfect measures of WTP_i. Despite these concerns, market data may well provide useful, though imperfect, evidence about economic values where such data exist.

However, for many goods and services, including many services from the natural environment, there are no markets. For various reasons, interventions involve more direct provision of such goods and services by governmental units. Nevertheless, market behavior may reveal something about the values of non-market goods and services. Expenditures to avoid environmental "bads" or repair the damages done by them reveal something about how much people would value being free of such bads. People may reveal their preferences for local environmental amenities when they buy and sell real property or accept lower incomes to live nearby. Travel and related costs may serve as proxies for market prices in estimating the demand for publicly provided recreational services.

One perhaps underutilized possible source of evidence about the values of non-market goods and services is voting behavior, especially voting in referenda. Votes for or against referenda involving commitments to pay taxes or other fees might provide credible evidence for estimating WTP_i, although questions do arise about such an interpretation. It is conceivable, for example, that people who themselves would favor a proposition (indicating that their WTP values exceeding the monetary commitments) might still vote against it if, for example, they felt that it was unfair to impose a tax on others. To call referenda "political markets" is perhaps stretching the concept of markets a bit. Nevertheless, the fact that voting for such referenda involves real commitments to pay gives voting clout as evidence about economic values. At least one study has explored this possibility (Deacon and Shapiro 1975).

Whereas traditionally only actual commitments of money in markets and perhaps in referenda have been acceptable as evidence of WTP_i, beginning 20

years ago, economic researchers began to introduce responses to CV survey questions as potentially valid evidence. When any new idea is introduced into a traditional discipline like economics, it will be subjected to considerable scrutiny. In the case of CV, the theory of public goods points to incentives that may lead people to intentionally state misleading values where environmental quality is to be valued, especially if real commitments of money are not required. This has led to the supposition that responses might bias contingent values away from WTP_i. From a psychological perspective (e.g., Kahneman and Knetsch (1992)), speculation has focused on whether people will have trouble constructing accurate estimates of WTP_i in response to survey questions.

Questioning of new methods of measurement is healthy in science, but progress comes from empirical work and not simply speculating about possible problems. The accuracy of any valuation method must be assessed in terms of how close it comes to measuring the ideal. If WTP_i were observable, then there would be no problem. One would simply observe it. Given that WTP_i is not observable, more complex criteria and "rules of evidence" are needed to assess accuracy. Continuing to borrow from psychology, accuracy in measurement depends on the reliability and validity of the data and analyses that are used in measurement.

Reliability and Validity

These concepts are clearly laid out by Mitchell and Carson (1989, pp.120-125). A measurement technique is unreliable if random errors in measurement reach unacceptable levels. Stated differently, measures are more reliable the less "noisy" are the data. Validity involves systematic errors in measurement. As Mitchell and Carson (1989, 190) have pointed out, "The validity of a measure is the degree to which it measures the theoretical construct under investigation." In the current context, the "theoretical construct under investigation" is WTP_i. A CV measure (or a measure derived using market evaluation or any other method) is "biased," and hence invalid, if it tends to

depart in a systematic way from WTP_i.

The economist is interested in reliability and validity on a different level from the psychologist. When psychologists use these terms they are primarily interested in accuracy at the level of the individual human being. Think about IQ testing, for example. While psychologists interested in intelligence may have some interest in the average IQ of aggregates of people, they are primarily concerned about measuring intelligence of individuals. On the other hand, economists seeking to evaluate changes in economic welfare are interested in the accuracy of average values for populations and aggregate over those populations. A considerable amount of unreliability in observed WTP is tolerable, provided bias is non-existent or at least within tolerable bounds. Accurate estimates of mean WTP_i can be made in the face of unreliability by simply increasing sample size. Reliability issues are not totally dissipated by this change in emphasis. For example, the reliability of mean estimated WTP over time may still be an issue. Unreliability in observed values at the individual level can lead to biased estimates of higher moments. Nevertheless, research on CV and other valuation methods focuses primarily on the validity of estimated average WTP over appropriately defined aggregates of individuals.

Current practice in CV requires that we raise one other issue relating to the concept of validity. Application of CV, like any empirical undertaking, involves numerous issues that must be resolved based on the judgement of the researcher. In damage assessments, and to a substantial degree in benefit-cost applications as well, researchers today will often resolve these issues by choosing the "conservative" course. That is, in exercising judgement, they make choices that will, if anything, lead to lower estimates for mean WTP_i. If, as some fear, CV studies tend to overestimate mean WTP_i, then conservatism, at least within limits, may pull estimated values toward the theoretical value. However, if, as others believe, CV tends to provided accurate measures of mean WTP_i, conservatism must be viewed as an expedient departure from normal scientific practice. To arrive at estimates of average WTP_i that are more easily defended in court or the policy arena, researchers make conservative

choices in study design and execution, thus intentionally introducing possible sources of downward bias. The target is no longer WTP_i, but an underestimate of it. Whatever the practical merits of conservatism, the ultimate scientific goal of CV studies is still to estimate aggregate WTP_i for the population in question. Conservatism in measurement must be viewed as an attempt to arrive at lower bounds for the desired values. In the long-run, refinements will hopefully lead to estimates that approach mean and aggregate WTP_i from below.

Because mean WTP_i is in principle unobservable, inferences about validity must be based on indirect evidence rather than direct comparisons of estimated WTP to WTP_i. Such indirect evidence may relate to the content validity, the construct validity, or the criterion validity of a measure.

Content Validity

Content validity has to do with whether the design and execution of a study are conducive to the revelation of WTP_i. In other words, assessing content validity involves examining the "content" of the study procedures. For CV studies, assessing content validity involves four steps. First, the study design must be compared to the economic theory underlying WTP_i. Second, the extent to which the study communicates effectively to the relevant population must be evaluated. These first two steps might be summarized by saying that a valid CV study must deal with both Homo economicus and Homo sapiens in ways that support WTP_i estimation. Third, whether various facets of study execution were adequate must be assessed. Here, such matters as sampling and response rates are considered. Fourth, the econometrics used to estimate mean and aggregate WTP_i and other statistics are examined. We now consider each of these aspects of content validity in more detail.

In dealing with Homo economicus, the scenario of a valid study provides the context and information that would lead a theoretical consumer to reveal his or her WTP_i. Fischoff and Furby (1988) provide a useful framework for developing a CV scenario with content validity. (Also see Bishop, Champ and Mullarkey 1995). The premise of Fischoff and Furby's framework is that a CV

exercise should fulfill the requirements of a satisfactory transaction. They define a satisfactory transaction as one "involving individuals who are fully informed, uncoerced, and able to identify their own best interests" (Fischoff and Furby 1988, 148). Three aspects of the transaction must be adequately defined and understood by participants: the good, the payment, and the marketplace. All three can affect the value an individual places on a good.

The good. "Goods may be thought of as bundles of *attributes*, representing outcomes of accepting the transaction that might be valued either positively or negatively" (Fischoff and Furby 1988, 153). The researcher must determine which attributes affect the value an individual places on a good, a particularly challenging task if the individual is not familiar with the good prior to receiving the survey. Fischoff and Furby also mention the importance of specifying the reference and target levels of the good. Borrowing from benefit-cost terminology, survey respondents need to know the level of provision of the environmental attributes "without" and "with" the intervention.

The payment. In most CV studies, the hypothetical payment is made in dollars. However, as use of CV expands to other contexts, alternative forms of payment such as labor hours are also being used (Swallow and Wondyalew 1994). Respondents need to be clear on the numeraire, whatever it is. The direction of payment must also be made clear. Will the individual pay for the good or will the individual be compensated? Another important aspect is the mechanism through which the payment will be made. Commonly used "payment vehicles" include income taxes, property taxes, sales taxes, entrance fees, changes in market prices of goods and services, and donations to special funds.

The marketplace or context for valuation. The third aspect of the CV scenario that must be specified is what Fischoff and Furby term the "marketplace." Since a market per se, as that term is normally understood, need not be involved, we prefer to think in terms of the "context" in which the transaction is to take place. The scenario may need to explain the extent of the market (who the other potential participants are), how and when the

environmental amenity might be provided, and the decision rule for the provision (e.g. majority vote, individual payment, etc.). Most researchers would agree that it is desirable to use an incentive compatible context in CV questions to avoid incentives for strategic responses.

Whether one takes the theoretical perspective of Homo economicus or the practical perspective of Homo sapiens, information on the good, the payment, and the context are central. Homo sapiens have certain additional needs that must be attended to. It is not difficult to imagine a study that covers all the theoretical aspects well but fails to effectively communicate the terms of the proposed "transaction" to the individuals whose values are to be estimated. Thus, Mitchell and Carson (1989, 192) recommended asking the following questions when assessing the content validity of a CV scenario:

> Does the description of the good and how it is to be paid for appear to be unambiguous? Is it likely to be meaningful to the respondents? Is there anything in the scenario that might suggest to some respondents that the good would not be paid for? Are the property right and the market for the good defined in such a way that the respondents will accept the WTP format as plausible? Does the scenario appear to force reluctant respondents to come up with WTP amounts?

CV researchers are increasingly recognizing that dealing with Homo sapiens is not a trivial problem. The result is increasing emphasis on "qualitative research" to support CV survey design. Focus groups, verbal protocols, observed one-on-one interviews, pretests, and pilot studies can be used to determine how information and questions can be most understandably presented.

While including qualitative research procedures in CV studies is commendable, their limitations in establishing the content validity of CV studies must be recognized. Focus groups, verbal protocols, observed one-on-one interviews, and other such procedures involve small samples which may not be representative. Furthermore, standard methods for conducting such procedures and for reporting results do not exist, at least at present. Typically, the outsider who is attempting to assess the content validity of a completed study is left with little more than a statement to the effect that a certain number of focus groups were conducted. There is always the danger that such exercises were conducted in such a way that points of confusion and

misunderstanding did not surface. Research to standardize procedures may be helpful. Pretests and pilot tests are more "quantitative" and hence more amenable to standardized procedures, but may not help identify more fundamental design flaws. For example, responses to pretests and pilots may not uncover ways in which respondents misunderstood the CV scenario. Inviting and carefully studying respondent's verbatim comments on the survey as well as including follow-up questions designed to probe for understanding may be helpful.

Clearly, qualitative research to enhance content validity is a very important part of any CV study, particularly studies of non-use values, where respondents may lack intimate familiarity with the environmental resources being valued. Much more research is needed to build consensus on proper procedures for conducting this phase of instrument development and reporting results. In the meantime, though they lack fully standardized methods, such procedures enhance content validity.

Finally, a study that is inadequate in its econometrics would be of questionable content validity. This aspect is noted for completeness, but little more needs to be said here. Economists normally have the training to do well in this regard. To the extent that CV studies involve competent applications of econometric methods, they will have higher content validity, all else equal.

In the end, content validity cannot be proven in some objective sense. Rather, it is up to the researcher to demonstrate to his or her peers that the study is designed and executed to be as conducive as possible to revelation of WTP_i. To be sure, some evidence can be accumulated that may support the content validity of a study. Questions can be added to the survey to generate data relevant to the assessment of content validity, for example. Results from focus groups and other qualitative procedures can be called upon for support as well. Content validity, nevertheless, is ultimately a matter of professional judgment.

Actual CV studies will display differing degrees of content validity.

That a given study has some apparent procedural flaws will not normally be grounds for totally dismissing its results. Still, the more such flaws are identified and the more serious they are judged to be by peer reviewers, the less credible are the estimates of value and more tenuous and tentative are any policy, damage estimates, and methodological conclusions derived from them. This is a point which has not been adequately recognized in the current literature. Instead, it is not unusual for authors to attempt to draw rather sweeping conclusions from results that many would judge as having limited content validity. While our primary purpose is to consider broad principles for assessing the validity of CV studies and not to give detailed reviews of individual studies, a couple of examples of studies from the current literature may help to clarify the sorts of problems we are concerned about.

Much has been made of the so-called embedding problem based primarily on the paper by Kahneman and Knetsch (1992). Drawing upon studies reported in their paper, Kahneman and Knetsch (1992, 58) concluded that embedding effects are "perhaps the most serious shortcoming of CVM [the contingent valuation method] . . ." Yet we have serious concerns about the content validity of the studies on which this conclusion is based. Consider their study to value improvements in the environment, disaster preparedness, and medical personnel and equipment at various levels of embeddedness. At all three levels of embeddedness, study participants were asked to value vaguely defined products under vaguely defined conditions. The reference levels of services were not defined at all and the target levels were merely described as "improvements." Nothing was said about which particular attributes would be improved, about the physical locations of the changes, about the timing of the changes, or other aspects. Instead of specific details about the terms of the proposed transaction, respondents were left with vague references to taxes, prices, and user fees to be placed in some undefined "specific fund." Given such weaknesses in content validity, how can generalizations about the shortcomings of CV be made?

Or, consider a second study. Based on several different CV treatments

involving logging of wilderness areas in the western U.S., Diamond et al. (1992, 15) concluded that, in general, "whatever contingent valuation surveys are measuring, they are not measuring consumers' preferences for environmental amenities." Though not so sweeping in his conclusions, McFadden (1994) used the same data set as a basis for voicing grave concerns about the usefulness of the CV method. Does this data set have sufficient content validity to support such a far-reaching conclusion? Respondents were told that seven or eight or nine wilderness areas, depending on the treatment, would be logged somewhere, but the location was not mentioned except that at least one would be located somewhere in the respondent's home state (Colorado, Wyoming, Montana, or Idaho). Via telephone interviews, respondents were asked to value one or more specified wilderness areas, again depending on the treatment, but these areas were only roughly described in terms of size and location, and little was said about how and when the logging would be conducted or what effects it might have on environmental parameters. Given all the vagueness in the information provided, respondents may be forgiven for expressing values that were not consistent with prior expectations of the researchers. It is not surprising, for example, that respondents missed the subtleties of whether seven or eight or nine other wilderness areas were already slated for logging. In addition, most CV researchers would shy away from a telephone survey for a study involving an issue of this complexity. Qualitative research might have addressed these concerns, perhaps even demonstrating that the telephone survey was a satisfactory approach, but no such enquiry is mentioned in the written materials from the study. Given that the content validity of the Diamond et al. study is highly questionable, then surely sweeping conclusions about the efficacy of the CV method as a whole based on the study results are equally questionable.

Construct Validity

Construct validity deals with the degree to which the measure under scrutiny (in our case observed WTP) relates to other measures as predicted by

theory. Mitchell and Carson (1989) discuss two forms of construct validity -- convergent and theoretical. Tests of convergent validity consider the relationship between the CV measure of the good's value and alternative measures of its value. For example, convergent validity could be assessed by comparing values estimated from a CV study to values estimated using a travel cost model or an hedonic pricing model. At least at present, this approach does not appear promising in the current context, since CV is widely considered the only method capable of estimating non-use values.

Theoretical validity is often assessed by considering the relationship between a CV measure and independent variables that theory indicates may be determinants of WTP_i. Assessing the theoretical validity of a measure may involve simple contingency table analyses. Or, more sophisticated multivariate regression procedures are applied and coefficients on potentially important independent variables are scrutinized for appropriate signs, statistical significance, and relative magnitudes.

Diamond et al. (1992), among others, have recently advocated a different form of theoretical validity test. They advocate testing theory-based hypotheses about relationships between two or more contingent values. For example, one would expect levels of WTP_i to be higher the more of an environmental amenity is provided or the larger is the environmental insult that is avoided. CV estimates of values should, if they are measuring WTP_i for different levels of provision that affect consumer welfare, exhibit relative magnitudes positively related to relative levels of provision. Tests of hypotheses about expected variations in estimated values with respect to changes in the size of environmental improvements or insults have come to be known as "scope tests." Within the taxonomy being followed here, scope tests are theoretical validity tests. Theory also tells us that values should be sensitive to the availability of complements and substitutes. Hence, one might construct CV exercises to test hypotheses about such effects. Or, transitivity is one of the core assumptions about preferences and might be used to develop construct validity tests for contingent values.

A fundamental question follows: What if the results from a CV application fail to pass a theoretical validity test? Suppose, for example, that a valuation equation fails to produce a positive relationship between observed WTP and income of respondents. Should the whole study be thrown out as invalid? Surely not. Failure to get a positive coefficient on income could even be consistent with theory. The value of the environmental amenity in question may simply be insensitive to income.

Now suppose that the results fail a scope test or some other test involving contingent value comparisons. Failure of a scope test would call for even more careful scrutiny of the content validity of the study. One would want to ask, for example, whether the qualitative phases of the research succeeded in identifying dimensions of the environmental amenity in question that mattered to potential respondents. The matter of whether the survey instrument communicated well would need to be pursued with extra vigor. Careful attention to statistical issues would also be warranted. Values may be inherently highly variable, and failure of a scope test might simply reflect the noisiness of the data. In the end, CV studies that fail scope tests will have less credibility that those that pass.

Suppose a study attempts to demonstrate sensitivity of values to availability of complements and substitutes or tests for transitivity and fails. Again this may indicate flaws in study design or statistical noise. Or, more fundamental problems may be indicated. As we have already noted, the economic theory on which such tests are based is highly abstract and stylized. In other words, reality is some unknown mixture of what theory models and what we termed "other things." Failure of responses to a CV survey to conform to theory may indicate that "other things" (phenomena not considered in the theory) are having a strong influence. What these "other things" are and whether they are somehow biasing results will usually not be very clear.

The overall conclusion would be that CV studies that display strong theoretical validity ought to be considered superior to those that display weaknesses in these respects or do not include theoretical validity tests in

their study design at all. This is so because studies with strong theoretical validity are more likely to be indicative of true values as defined in theory. Suspected weaknesses identified during theoretical validity testing may indicate flaws in study design that were not detected when the content validity of the study was assessed. Or, they may arise because unknown factors outside the theory used to define WTP_i are influencing results.

It seems likely that there will be wide variations in how well CV studies perform in theoretical validity testing. We propose that theoretical validity tests be divided into two categories, which we shall term rudimentary and advanced tests. Through regression analysis, contingency tables, or other such procedures, the rudimentary tests would explore whether expected theoretical relationships can be found between contingent values, on the one hand, and socioeconomic, behavioral, and attitudinal variables, on the other. Independent variables could include income, socioeconomic characteristics, and attitudes relative to the resources in question and the environmental issues more generally. Having previously visited the site, membership in environmental organizations, participation in recreational activities related to the environmental resources in question, past volunteer activities, and other such variables could also be considered in such analyses. Rudimentary tests should be feasible using data from a relatively simple survey of a single sample. They are "rudimentary" in the sense that they are relatively cheap and straightforward to perform and probably ought to be a part of all CV studies. It should be recognized from the beginning that results from a study that lacks the money and/or time to gather and analyze information on at least some such variables will likely not have much credibility.

Advanced construct validity tests involve comparisons between contingent values. Theoretical expectations relative to scope, variations in substitutes and complements, and other relationships between values are tested. Such tests may be beyond the resources available to many studies not only because they will involve more sophistication in survey design and execution, but also because they may involve costly split-sample designs.

We propose that studies be categorized into a three-level hierarchy expressing increasing degrees of construct validity. At the lowest level would be studies that either have not included any construct validity tests or have failed to pass rudimentary tests. Such studies might typically have had low budgets and/or severe time constraints and this may have limited the amount of qualitative research that could be conducted, thus limiting the content validity of the study as well. Such studies may be useful for scientific purposes or as exercises involving training of students, but should be used in policy analysis and litigation only with the heaviest caveats. The second level of the hierarchy would involve studies that have achieved a fair amount of success in the rudimentary tests, but that either do not have the budget to support advanced testing or have not succeeded in passing advanced tests. Second-level studies may be usable in cost-benefit analyses, since normally the goal of such analyses is simply to determining whether the benefits of an intervention exceed the costs. Of course, unless benefits exceed costs by a fairly wide margin or vice versa, potential imprecision in second level studies may mean that the issue of whether benefits exceed costs remains open. Second level studies may be less applicable in litigation, where relatively precise estimates of value are needed to assess damages, but they may still be useful in preliminary damage assessments. Third level studies are studies that have conducted and achieved substantial success in passing advanced tests. Provided that such studies are judged to have a high degree of content validity as well, they would have the highest level of credibility for benefit-cost analysis and litigation.

The other issue needing attention in this section is the extent to which success or failure in construct validity tests implies that CV itself is a valid or invalid procedure. Notice that the focus has changed substantially. Up to now we have been concerned about the validity of individual applications. Now the ability of the CV method in general to estimate values of WTP_i is the issue. Can construct validity testing help to address this much broader question?

The difficulties created by our inability to observe true values become painfully apparent at this point. While failure to pass construct validity tests does, as we have just seen, create some degree of doubt about individual studies, the problems thus identified may be peculiar to the study in question. Perhaps design flaws in the individual study are to blame. Such flaws may have been identified when content validity was considered, but they may also have been hidden. Perhaps the "other things" that theory ignores are more potent or play themselves out in some unusual way in the study in question. For these reasons, the temptation to draw sweeping conclusions about the validity of the CV method from one or only a few studies should be resisted. Only if a large number of seemingly high-quality CV studies consistently fail rudimentary and/or advanced tests, would a negative conclusion regarding the overall validity of the CV method in general be warranted. Likewise, if CV studies consistently pass such tests, then the validity of the method would be supported. In the intermediate case where CV studies sometimes pass such tests and sometimes fail them, a very different conclusion would be warranted. Assuming that the studies in question have a high degree of content validity, such an outcome would imply that intervening factors are sometimes, but not always, interfering in the successful application of the method. Research would be called for to identify what those intervening factors are. New content and construct validity criteria would follow. Ultimately, it may be possible to develop criteria for judging where CV can be applied with good prospects for success and where it cannot. In the meantime, the conclusion regarding the overall validity of the CV method would be that it is capable of producing results that pass validity tests under at least some circumstances. We will return to the top topic of overall validity after considering the third of the Three Cs.

Criterion Validity

To assess criterion validity, Mitchell and Carson (1989, 192) point out that "it is necessary to have in hand a criterion which is unequivocally closer

to the theoretical construct [*WTP*, or its mean value over some population] than the measure whose validity is being assessed [the CV-based measure of value]. " The closer the contingent value is to the criterion, the more valid it is judged to be.

Given the credibility that preference expressions in markets have as indicators of true values, actual market prices would be ideal criteria to use in evaluating contingent values; however, because such market prices are rare in the environmental area and especially for non-use values, so-called "simulated market" values are perhaps a more promising alternative for judging the criterion validity of contingent values. Simulated markets involve creating situations in the field or laboratory where subjects have the opportunity to actually pay for the good or service or receive compensation for it. CV is used to estimate the value of the same good or service.

Simulated markets differ from real markets in two ways. First, each individual may be involved in only one transaction. Second, the mechanism by which the price is determined may seem somewhat artificial to participants.

Bohrnstedt (1983) provides a useful distinction between two kinds of criterion validity: predictive validity and concurrent validity. Applied in the context of CV, predictive validity might be assessed by asking subjects a CV question at one point in time and later giving the same individual a chance to actually purchase the same good (see, for example, Kealy et al. 1990). For concurrent validity, the measure and the criterion against which it is to be assessed are measured contemporaneously. This could involve a split-sample design where randomly assigned subjects participate in either a CV exercise or a simulated market in which the same good can actually be bought or sold (see, for example, Bishop and Heberlein 1979).

Simulated market experiments have been conducted for both use values (Bohm 1972; Bishop and Heberlein 1979; Dickie et al. 1987; Coursey et al. 1987; Kealy et al. 1990; Bishop, et al. 1993; Neill et al. 1994) and total values where non-use values are likely to be a substantial component (Boyce et al. 1989; Kealy et al. 1990; Duffield and Patterson 1992; Seip and Strand 1992; Champ et

al. 1994). It is tempting to launch into a discussion of these studies, for they provide many interesting features and results. Space and time preclude doing justice to such an exercise. It must suffice to say that the evidence is somewhat mixed, with several studies providing support for the criterion validity of CV, especially for use values. Results from simulated market experiments that have focused on total values (including non-use values) are even farther from being conclusive. Furthermore, for both use and total value studies, problems arise that are not unlike those already mentioned for construct validity. Circumstances, commodities to be valued, procedures applied, and other details vary so greatly from one study to the next, that it is difficult to generalize from the studies that exist to the method in general. For example, in Kealy et al. (1990), Duffield and Patterson (1992), Seip and Strand (1992), and Champ et al. (1994) there is substantial evidence that donation vehicles lead to overestimation of total values, but little evidence exists from simulated markets about how other payment vehicles perform. Clearly, the challenges of conducting a simulated market to estimate non-use values with a high degree of content validity are formidable.

An alternative with some promise would be to use CV to predict the share of the population that would vote in favor of propositions in actual referenda. The potential usefulness of referenda as indicators of true values was discussed above. It is worth adding that the advantages of the referendum format for CV non-use value studies are well established in the literature (Mitchell and Carson 1989; Hoehn and Randall 1987). At least one criterion validity study using this approach has already been conducted (Carson, Hanemann, and Mitchell 1986). They were able to successfully forecast the outcome of an actual referendum in California using CV, a result that supports the validity of the CV method. Though additional studies of this kind should be pursued, their disadvantage relative to simulated markets should be recognized. True values, as defined here, are monetary values. Voting in real referenda will not produce direct evidence to estimate true values that can serve as a standard of comparison for contingent values. Referenda can only

produce as criteria the percentages of citizens voting for and against ballot propositions. Econometric models of CV responses would then be used to predict the vote based on survey responses. The null hypothesis to be tested is that the CV-based predicted vote equals the actual vote. Results would obviously be relevant to understanding the validity of CV, but would not be fully equivalent to comparing monetary values.

Some Final Thoughts on the Overall Validity of the Method

We might sum up the evidence on the overall validity of the CV method by saying that, in both construct and criterion validity testing, CV seems to perform well in some studies and not so well in others. Carson et al. (forthcoming) suggest a new way to try to come to grips with variety of results that are now available. They located a total of 83 studies that contained both revealed preference and contingent values for the same interventions. Some of the revealed preference results came from criterion validity studies where simulated market or actual market comparisons were feasible. Other studies involved comparisons of contingent values with hedonic price, travel-cost, and averting expenditure studies and would probably be best considered convergent validity comparisons. All the studies consulted involved WTP. Carson et al. calculated a total of 616 possible ratios of contingent values to revealed preferences from these studies. These ratios were analyzed in several different ways, but the simplest approach, which took the ratios as a data set containing 616 observations, is fairly typical. The ratios averaged 0.89 with a 95 percent confidence interval of [0.81-0.96] and a median of 0.75. The Spearman rank correlation coefficient for contingent values and associated revealed preference values was 0.78.

Caution is warranted in interpreting these results. Only where use values dominate are revealed preference values available to calculate such ratios. Non-use values may present special difficulties that were not present for use values. Also, such averages do not guarantee that any particular CV study, no matter how well or poorly is was conducted and regardless of specific

circumstances, will get a value comparable to what would have been obtained had a revealed-preference study been done. Granting all that, both the closeness of this ratio to unity and the tightness of the confidence interval are impressive. It would appear that for use values at least, contingent valuation studies tend to get values that are comparable in magnitude to revealed preference values. And, if people can answer CV use value questions in ways that tend to be close to the value of the same resource from revealed-preference based studies, then this will tend to support those who argue that they can answer well-designed CV total value questions in ways that reflect their underlying economic preferences.

Further favorable evidence is provided by noting that both use and total value CV studies frequently pass scope tests, as Richard Carson's paper in this volume shows. While not all studies pass rudimentary and advanced tests, enough do to conclude that CV is capable of producing results of sufficient validity to be useful in policy analysis and damage assessment.

Summary

The debate over CV is a debate that can only be resolved with empirical evidence. Too often, it seems to us, the critics of CV simply speculate about what might go wrong in CV studies and then assume that those things actually do go wrong. On the other hand, CV practitioners have probably been far too willing to accept the results of their studies at face value. Carrying on the debate at this level will continue to bear little fruit.

In order to consider the empirical evidence systematically and objectively, the ground rules must be determined in advance. What is needed is a theory of measurement for CV that can guide the empirical research. To that end, we have attempted in this paper to expand on a theme that already exists in the literature on CV. The approach builds on the theory of psychological testing. It is based on the premise that true values are unobservable, so that validity must be evaluated indirectly, through drawing of inferences from a number of directions. In psychology, reliability must be carefully considered

along with validity. If the data are too noisy, it will be impossible to detect biases in observations at the individual level even if the noise is simply random. Reliability is somewhat important for CV as well, but it takes on less importance because, in economics, the goal is normally to measure values for a population, and not for the individual. Individual values could be quite noisy, yet average values could be unbiased and accuracy would simply be a matter of sample size.

The Three C's--content validity, construct validity, and criterion validity--seem to us to fit the problem of evaluating the validity of CV rather well. The first step toward achieving a valid CV study--if that is possible--is to design and execute the study properly, and the goal of content validity assessment is to evaluate the study procedures. Since the goal is to measure the theoretical WTP, a well-designed study will show a close correspondence between the economic theory which is used to define the true value and the structure and wording of the CV question. Stated differently, everything that neoclassical consumers would need to formulate and reveal their true values must be present in the CV exercise. Furthermore, since normal human beings, and not theoretical consumers, will be asked to complete the survey, care must be taken to design the survey to be effective in communicating with real people. There are numerous hidden pitfalls here that may require a substantial amount of qualitative research to overcome. Furthermore, a study with high content validity will have carefully executed all phases of the survey, from population definition and sampling through coding and entry of the responses. Finally, a study with high content validity will have applied appropriate econometric procedures to arrive at values and related statistics.

For non-use valuation studies, construct validity assessment will mainly involve testing theoretically motivated prior expectation about relationships between the contingent value estimate or estimates in question and other, non-value variables (rudimentary tests) or other contingent values (advanced tests). Studies that fail to show expected relationships in rudimentary tests will have the least credibility. Such studies fall at the bottom of our

hierarchy of construct validity. At the top of the hierarchy will be studies that pass a battery of rudimentary and advanced tests of theoretical validity.

Theoretical validity testing can shed some light the broader question of whether the CV method itself is valid. Because individual studies vary widely in their subject matter and content and construct validity, the weight of evidence from many studies will be required in order to judge the overall validity of CV. The temptation to reach broad conclusions on the basis of one or a few studies should be resisted. If, across a wide range of studies, it had turned out that CV consistently fails theoretical validity tests, then we would conclude that CV, at least as it has been applied in the past, is a scientific dead-end. However, as we have seen, this has not been the case. Instead, while not all studies have performed well under theoretical validity testing, many have performed quite well in both rudimentary tests and scope tests.

Criterion validity tests may be capable of casting further light on the overall validity of the method. Here, simulated market experiments--while they involve significant challenges in design and implementation--may yet turn out to be a productive avenue. Further efforts to find circumstances where simulated markets are feasible could produce rich dividends in judging the potential of the method. Here again, individual studies are unlikely to be definitive. Also, care will always be required in extrapolating conclusions from such simulated markets to specific field applications of CV. Content and construct validity will continue to play a central role in evaluating the validity of individual CV studies, with criterion validity studies hopefully helping to refine the criteria applied to individual studies.

References

Bishop, R.C., P.A. Champ, D.J. Mullarkey. 1995. "Contingent Valuation." *Handbook of Environmental Economics*. Ed. Daniel W. Bromley. Oxford: Blackwell Publishers, pp. 629-654.

Bishop, R.C. and T.A. Heberlein. 1979. "Measuring Values of Extramarket Goods: Are Indirect Measures Biased?" *American Journal of Agricultural Economics* 61(5):926-930.

Bishop, R.C. and M.P. Welsh. 1992. "Existence Values in Benefit-Cost Analysis and Damage Assessment." *Land Economics* 68(4):405-417.

Bishop, R.C., M.P. Welsh, and T.A. Heberlein. 1993. "Some Experimental Evidence on the Validity of Contingent Valuation." Unpublished paper, Department of Agricultural Economics, University of Wisconsin--Madison.

Bohm, P. 1972. "Estimating Demand for Public Goods: An Experiment." *European Economic Review* 3:111-130.

Bohrnstedt, G.W. 1983. "Measurement." *Handbook of Survey Research*. Eds. P.H. Rossi, J.D. Wright, and A.B Anderson. New York: Academic Press, Inc.

Boyce, R.R., T.C. Brown, G.D. McClelland, G.L. Peterson, and W.D. Schulze. 1989. "Experimental Evidence of Existence Value in Payment and Compensation Contexts." Paper presented at the joint meetings of the Western Committee on the Benefits and Costs of Natural Resource Planning (W133) and the Western Regional Science Association, San Diego, CA.

Boyle, K.J., and R.C. Bishop. 1987. "Valuing Wildlife in Benefit-Cost Analyses: A Case Study Involving Endangered Species." *Water Resources Research* 23(5):943-950.

Carson, R.T., N.E. Flores, K.M. Martin, and J.L. Wright. Forthcoming. "Contingent Valuation and Revealed Preferences Methodologies: Comparing the Estimates for Quasi-Public Goods." *Land Economics*.

Carson, R.T., W. Michael Hanemann, and Robert C. Mitchell. 1986. "Determining the Demand for Public Goods by Simulating Referendums at Different Tax Prices." Manuscript, University of California, San Diego.

Champ, P.A., R.C. Bishop, T.C. Brown and D.W. McCollum. 1994. "Some Evidence Concerning the Validity of Contingent Valuation: Preliminary Results of an Experiment." Ed. R.C. Ready, W-133 Seventh Interim Report. Western Regional Research Publication, Lexington, KY.

Coursey, D.L., J.J. Hovis, and W.D. Schulze. 1987. "The Disparity between Willingness to Accept and Willingness to Pay Measures of Value." *Quarterly Journal of Economics* 102:679-90.

Deacon, R. and P. Shapiro. "Private Preference for Collective Goods Revealed Through Voting in Referenda." *American Economic Review* 65(5):943-955.

Diamond, P.A., J.A. Hausman, F.R. Johnson, R.W. Dunford, K.J. Boyle, S.P. Hudson, and K.N. Wilson. 1992. "Measuring Natural Resource Damages with Contingent Valuation: Tests of Validity and Reliability." Paper presented at the Cambridge Economics, Inc. symposium titled "Contingent Valuation: A Critical Assessment", Washington, D.C., April 2 and 3.

Dickie, M., A. Fisher, and S. Gerking. 1987. "Market Transactions and Hypothetical Demand Data: A Comparative Study." *Journal of the American Statistical Association* 82:69-75.

Duffield, J.W., and D.A. Patterson. 1992. "Field Testing Existence Values: Comparison of Hypothetical and Cash Transaction Values." Paper presented at the joint meetings of the Western Committee on the Benefits and Costs of Natural Resource Planning (W133) and the Western Regional Science Association, South Lake Tahoe, Nevada.

Fischoff, B., and L. Furby. 1988. "Measuring Values: A Conceptual Framework for Interpreting Transactions with Special Reference to Contingent Valuation of Visibility." *Journal of Risk and Uncertainty* 1:147-184.

Freeman III, M. 1993. *The Measurement of Environmental and Resource Values*. Washington, D.C.: Resources for the Future.

Hoehn, J.P., and A. Randall. 1987. "A Satisfactory Benefit Cost Indicator from Contingent Valuation." *Journal of Environmental Economics and Management* 14(3):226-47.

Kahneman, D. and J.L. Knetsch. 1992. "Valuing Public Goods: The Purchase of Moral Satisfaction." *Journal of Environmental Economics and Management* 22(1):57-70.

Kealy, M.J., M. Montgomery, J.F. Dovidio. 1990. "Reliability and Predictive Validity of Contingent Values: Does the Nature of the Good Matter?" *Journal of Environmental Economics and Management* 19:244-263.

Krutilla, J.V. "Conservation Reconsidered." *American Economic Review* 57:777-86.

Madariaga, B., and K.E. McConnell. 1987. "Exploring Existence Value." *Water Resources Research* 23(5):936-942.

McFadden, D. 1994. "Contingent Valuation and Social Choice." *American Journal of Agricultural Economics* 76(4):689-708.

Mitchell, R.C., and R.T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, D.C.: Resources for the Future.

Neill, H.R., R.G. Cummings, P.T. Ganderton, G.W. Harrison, and T. McGuckin. 1994. "Hypothetical Surveys and Economic Commitments." *Land Economics* 70:145-154.

Randall, A., and J.R. Stoll. 1983. "Existence Value in a Total Valuation Framework" in *Managing Air Quality and Scenic Resources at National Parks and Wilderness Areas*. Eds. R.D. Rowe and L.G. Chestnut. Boulder, CO: Westview Press.

Seip, K., and J. Strand. 1992. "Willingness to Pay for Environmental Goods in Norway: A Contingent Valuation Study with Real Payment." *Environmental and Resource Economics* 2(1):91-106.

Smith, V.K. 1987. "Nonuse Values in Benefit-Cost Analysis." *Southern Economic Journal* 51:19-26.

Swallow, B.M. and M. Wondyalew. 1994. "Evaluating Willingness to Contribute to a Local Public Good: Application of Contingent Valuation to Tsetse Control in Ethiopia." *Ecological Economics* 11:153-161.

Footnotes

1. Bishop is a Professor in the Department of Agricultural Economics and the Institute for Environmental Studies at the University of Wisconsin--Madison. Champ was formerly a Graduate Student in the Department of Agricultural Economics at the University of Wisconsin--Madison and is now an economist at the Rocky Mountain Forest and Range Experiment Station. Brown and McCollum are economists at the Rocky Mountain Forest and Range Experiment Station, Forest Service, United States Department of Agriculture. The research was supported by the Rocky Mountain Forest and Range Experiment Station, by the College of Agricultural and Life Sciences, University of Wisconsin--Madison, and by the Wisconsin Sea Grant Institute (under grants from the National Sea Grant College Program, National Oceanic and

Atmospheric Administration, U.S. Department of Commerce, and from the State of Wisconsin; Federal grant number NA90AA-D-SG469, project number R/PS-38).

2. We shall limit the discussion to interventions that affect the environment, although everything said is directly applicable to other types of interventions.

3. Care must be taken in this sort of discussion to avoid making theory into a straw man. Economists interested in household decision making, decision making under uncertainty, and other economic problems do develop much richer theoretical models than those typically drawn upon in conceptualizing WTP_t for purposes of applied welfare analysis. Perhaps richer theories will eventually yield new insights and testable hypotheses for valuation studies, but they have yet to do so. To date, the simpler models, which intentionally leave numerous potentially important other factors out of the analysis, have dominated discussions of WTP_t.

APPENDIX B:

ASSESSING THE CONTENT VALIDITY OF CONTINGENT VALUATION STUDIES

Assessing the Content Validity of Contingent Valuation Studies

Richard C. Bishop and Daniel W. McCollum*

Please address all correspondence to: Professor Richard C. Bishop
Department of Agricultural Economics
Taylor Hall, 427 Lorch Street
University of Wisconsin
Madison, WI 53706
(608) 262-8966
FAX: (608) 262-4376
rcbishop@calshp.cals.wisc.edu

* Professor, Department of Agricultural Economics, University of Wisconsin--Madison, and Economist, USDA, Forest Service, Rocky Mountain Forest and Range Experiment Station, Fort Collins, Colorado, respectively. Research supported by the Rocky Mountain Forest and Range Experiment Station and the College of Agricultural and Life Sciences at the University of Wisconsin--Madison. Thomas C. Brown, Patricia A. Champ, Robert K. Davis, John Loomis, and Norman Meade provided many helpful comments on earlier drafts. This paper also benefited greatly from comments of the members of W-133.

Assessing the Content Validity of Contingent Valuation Studies

Abstract: Content validity assessment involves evaluation study procedures. This paper proposes a set of content validity criteria for contingent valuation studies and a rating form for use in assessing how well studies were designed and executed. The form's goal is to help researchers design content valid studies and reviewers to conduct more systematic, balanced validity assessments.

Quoting Mitchell and Carson (p. 190), "The validity of a measure is the degree to which it measures the construct under investigation." In applied welfare economics, the construct is most often one of the Hicksian measures of economic value. Assessing the accuracy of consumer welfare measures is difficult because true Hicksian values are inherently unobservable. Hence estimated values cannot be compared directly with true values to judge the performance of measurement techniques (Bishop et al. 1994). This is the case whether the valuation technique in question is contingent valuation (CV) or one of the methods that attempts to infer values from revealed-preference data. Hence, less direct forms of evidence about the validity of valuation techniques are required.

The debate over CV, spawned in part by work surrounding the Exxon Valdez oil spill (Carson et al. 1992; Hausman 1993), is a debate over the validity of the method. Though encouraged by the adversarial context of natural resource damage assessment, this debate is symptomatic of a more serious and fundamental gap in applied welfare economics. Advocates of CV are proposing that survey evidence about economic values be accepted in an arena where revealed preference evidence has long dominated. Progress in considering this proposal is hampered by a lack of consensus among economists regarding criteria for judging the validity of welfare estimates, and this is true

whether revealed preference measures or CV measures are being considered. More succinctly stated, applied welfare economics lacks a theory of measurement. The goal of this paper is to work toward such a theory. While we focus on CV, we believe that our work has implications for economic measurement more generally. In point of fact, revealed preference measures deserve much more careful and systematic scrutiny than they have received in the past and consistent criteria for revealed preference and CV approaches--and other possible measurement tools--should be a long term goal.

Lacking an economic theory of measurement, CV researchers (e.g., Mitchell and Carson; Bishop et al. 1994, 1995) are turning to other disciplines that have struggled to assess the validity of empirical measures of unobservable constructs, particularly psychology. In psychometrics, the validity of a measure may relate to its content validity, construct validity, or criterion validity (American Psychological Association; Sundberg; Zeller and Carmine; Bohrnstedt). Each of these approaches "offers a different strategy for assessing the measure-construct relationship, and each is applicable to contingent valuation in one way or another." (Mitchell and Carson, p. 190) This paper focuses on content validity.

Content validity assessment involves evaluation of study design and execution.[1] Partly, it is guided by theory. Measured values will ultimately be interpreted as estimates of values as defined in theory. From a practical standpoint, this means that CV instruments materials must be designed in ways that would support revelation of true values by the consumer of economic theory. Hence, for a CV study to be fully content

valid, respondents must have incentives for true value revelation and enough information to make utility maximizing choices. Content validity assessment also asks whether CV study procedures were designed to interact effectively with potential respondents. It is not hard to imagine a study that is strongly linked to theory, yet fails to deal well with real people. Through experience, CV researchers have learned this is not a trivial problem. Finally, to be content valid, the survey, subsequent analysis, and presentation of results must be adequately executed. Here, attention is focused upon such topics as sampling, response rates, and econometric procedures.

Though most did not frame their work in terms of content validity, many writers on CV have addressed issues of study design and execution. The reference operating conditions of Cummings, Brookshire, and Schulze represented an early attempt to explicitly state some validity criteria for CV studies, including content validity criteria. We draw much from Mitchell and Carson, particularly in the area of CV survey design. Recent contributions to the literature on procedural issues include the report of the NOAA Panel on Contingent Valuation (U.S. Department of Commerce) and the recent paper by Hanemann.

Figure 1 will serve as the centerpiece for the paper. We propose it as a tool for systematically rating the content validity of CV studies. The rating form is our attempt to synthesize past literature that relates, in one way or another, to CV content validity and our own experience as researchers. Our purpose is <u>not</u> to attempt to set up ourselves or anyone else as ultimate authorities on the content validity of CV studies. Rather we

hope simply to make content validity assessment more systematic and explicit. All of us involved in CV research consider some studies to be stronger than others. Such judgements are based in part on how well such studies appear to have been designed and executed. The rating form is intended to help reviewers, in their various roles[2], to be methodical in evaluating CV study procedures and clearer about their reasons for judging those procedures to be strong or weak. Hopefully, Figure 1 will also help those who conduct CV studies to improve procedures by explicitly stating a set of standards that others will likely use to judge what they have done. The rating form can be viewed as a checklist of considerations that should be addressed in designing and executing studies that aspire to high content validity.

Figure 1 begins with 12 questions about the detailed study procedures. Points are to be assigned to the study under review depending on how well it did in addressing the issues raised under each question. After devoting some attention to preliminary matters in Section I of the paper, each of the detailed questions will be explained and justified in Section II. Following the 12 detailed questions, the rating form asks reviewers to summarize their evaluations of study procedures by adding up the points on the individual items, by explicitly stating any concerns they have that were not covered by the detailed questions, and by rating overall study procedures on a five item scale ranging from excellent to unacceptable. Issues relating to the adding up of points and the overall rating are dealt with in the third section. At the end, the major points of the paper are summarized and some final thoughts offered.

Obviously, judgements about the validity of any study will depend on more than its content validity. Important additional evidence will come from subjecting study results to hypothesis testing based on theoretical expectations (i.e., construct validity testing). How CV has fared in laboratory and field experiments and other efforts to test its criterion validity will also be relevant. Such broader issues are beyond the scope of content validity assessment and are dealt with elsewhere (Mitchell and Carson; Bishop et al. 1994, 1995).

I. PRELIMINARY ISSUES

Professional Judgment and the Burden of Proof

As a point of departure, let us suppose that a proposed "intervention" in the economy affects environmental attributes (or other economic paramenters) relevant to some human population. Such an intervention could take the form of a public project, an alteration in environmental regulations, or a new policy that somehow affects the environment. The intervention could also take the form of an accidental or intentional environmental insult such as an oil spill or emission of air pollutants. Suppose further that a CV study has been conducted to estimate the values that members of the affected human population place on enjoying the positive effects of the intervention or avoiding its negative effects. Content validity of such a study would be conducted in the context of two over-arching principles.

First, content validity assessment is inherently a matter of professional judgement.

Because there is less than complete consensus about CV procedures, study designers and reviewers must inevitably fall back on personal judgement. At least for the time being, whether procedures are flawed and the seriousness of any flaws remains a matter for individual reviewers to judge based on their interpretations of their own work, if any, and the larger literature. It follows immediately that the cogency of the conclusions from a content validity assessment depend directly on the credentials of the reviewer.

Second, the burden of proof regarding the validity of study's procedures rests with the researchers who designed and executed the study. Replication is problematical in survey research. Furthermore, CV procedures are far from standardized. As a result, content validity assessment involves an evaluation of the detailed study procedures. Researchers must make the case for the content validity of their studies.

Criteria and Points

As we envision it, content validity assessment will require reviewers to consider how well or poorly the study did in addressing a list of procedural issues. Our rating form (Figure 1) is designed to capture the major issues. We do not expect the issues raised there to be particularly controversial. However, we propose that reviewers answer each of the 12 detailed questions with a numerical scores. The maximum attainable numerical scores we assigned to the different dimensions are likely to be more controversial. The current state of the art in CV leaves a great deal of room for debate on the relative importance of different aspects of study design and execution. To deal

with this problem, Figure 1 is amenable to whatever weights a particular researcher or reviewer deems appropriate.

As the review of any particular study proceeds, potential flaws in procedures will almost certainly be identified. Such potential flaws will often not be judged fatal, though the possibility of fatal flaws exists and will be dealt with later. Content validity assessment often involves the identification of potential flaws. That is, in the course of the assessment, doubts arise about whether procedures followed might have led to biased results. In more colloquial terms, content validity assessment involves a search for what are commonly termed "red flags." The more such red flags pop up during evaluation of a study, the less valid it will be judged to be. Our scoring system is designed to, in a sense, count red flags, or rather the lack of them.

On any particular item in the form, some studies may easily receive full credit simply because an issue did not arise in that particular case. Other studies may lose points for having neglected to one degree or another the issue or issues highlighted in the question. Under particularly difficult circumstances, a study may receive a low score despite competent efforts to overcome a particularly knotty problem. This would simply reflect the difficult circumstances that are present in that particular case. It should be more difficult to establish the content validity of CV studies in some situations than in others.

Flaws may creep into CV studies through simple lack of foresight on the part of study designers. Furthermore, some flaws are knowingly accepted as compromises

required to achieve other goals. For example, in some situations, a referendum format or some other mechanism with theoretically strong incentive characteristics may be very implausible to potential respondents. One might adopt a donation payment vehicle in such situations, notwithstanding its theoretical inferiority. Despite the fact that the researcher made this compromise intentionally and after full consideration of the alternatives, the use of an incentive incompatible mechanism would reduce the content validity of the study (in our opinion!) and this should be recognized in the score assigned under the question that relates to incentive compatibility (Question 5, as discussed below).

II. THE DETAILED QUESTIONS

Having laid a foundation for the rating form, we now look at its detailed questions, Question 1 through 12 in Figure 1. In each case, we explain the nature of the issues raised, attempt to assess their importance, and suggest the number of points that we believe that particular question warrants.

(1) Was the theoretical true value clearly and correctly defined?

Study designers may strengthen the link between theory and the CV exercise--thus enhancing content validity--by carefully defining, in theoretical terms, what is to be measured. The simplest model of the consumer's choice problem where environmental quality matters will illustrate. Such a consumer would solve the problem:

$$\max U(X;Q) \text{ subject to } P'X \leq Y,$$

where X is a vector of conventional goods and services that can be purchased at exogenously determined prices P, Q is an exogenously determined vector conveying the status of environmental attributes affecting consumer welfare, Y is income, and U(.) is a "well-behaved" utility function. Assume that the only effect of the intervention in question is to alter the status of environmental attributes, let us say from Q' to Q".

Theory tells us that the maximum level of utility, arrived at by solving the choice problem just stated, can be expressed as an indirect utility function, V(P,Q,Y). Assuming that the Hicksian compensating welfare measure is relevant, the "theoretical true value" of this intervention to the consumer, which we shall symbolize by T, is defined by

$$V(P,Q',Y) = V(P,Q'',Y-T).$$

Now suppose a CV study is to be conducted to estimate the mean value of T for the policy-relevant population. The benefits of formally considering the theoretical true value are many, as even this simple model illustrates. For example, for respondents to arrive at their estimates of T, they would have to be "well informed" about how the intervention would affect relevant parameters of their choice problem. Respondents would not be well informed if information is unavailable to them that a theoretical consumer would find relevant in solving the utility maximization problem.

Definitions of value should not only be clear, they should be "correct." That is, the researcher should make the theory fit the problem at hand. Some studies will be able

to focus on effects of the intervention on environmental attributes alone, as we did in the model just presented. Other studies may have to deal with effects on prices, incomes, and other parameters as well. The timing of both effects and payments may affect true values. Where uncertainty of one kind or another is a potentially significant factor in the theoretical consumer's valuation problem. Designers of CV studies should carefully consider the definition of T applicable in their particular case. Formal theoretical modeling of the valuation problem never hurts. Writing out the equations may seem mundane, but can prove helpful in identifying gaps and flaws in the information and context that will ultimately be provided in the CV scenario. Clearly defining the theoretical true value appropriate in the particular application may help to successfully address issues under many of the later questions on the form, especially Questions 3, 4, 5, 11, and 12.

The rating form allows up to 5 points to be assigned to a study depending on how well it defined the true value or values it sought to measure.

(2) Were the environmental attributes relevant to potential subjects fully identified?

In the abstract world of theory, the environmental attributes affecting consumer welfare can be represented by including the vector Q in the direct and indirect utility functions. However, theory alone offers limited guidance regarding which actual attributes are relevant to real world study subjects and which are not. From the potentially large set of attributes of the environment that might be relevant in theory, a

subset that human respondents believe affects their welfare must be defined.

Introspection and casual observation on the part of the researchers help to formulate working hypotheses about which attributes might be relevant. For example, it seems likely that attributes affecting human health are important to people. However, once such obviously relevant attributes are identified, it may be necessary to use more formal, empirical methods to sort out which attributes matter. CV studies often employ focus groups for this purpose. Researchers may also observe one-on-one interviews with subjects from the pool of potential respondents. Such interviews and particularly debriefing session with subjects afterwards can help sort out the relevant attributes. Verbal protocols (Schkade and Payne) may be analyzed to further explore how respondents view the attributes. Such "qualitative research techniques," if competently applied, will enhance content validity.

We have allocated up to 10 points for this aspect. How many points to assign to a study will vary depending on the particular circumstances. Studies where respondent-relevant attributes are rather simple and obvious may earn the full 10 points after little or no qualitative research. Other interventions may have effects which are complex and less obviously relevant to people. In such cases, reviewers might assign fewer than 10 points in recognition of the inherent difficulty of the problem.

(3) Were the potential effects of the intervention on environmental attributes and other economic parameters adequately documented and communicated?

Following determination of the environmental attributes relevant to potential study

subjects, the next step in study design is to document how the intervention will affect those attributes. This is normally done by finding out what physical and biological scientists know (and do not know) about the effects of the intervention. Impacts on non-environmental parameters such as prices and incomes also need to be documented in cases where they could occur. The more thoroughly such effects were investigated and documented, the higher should be the score on this item.

Once potential effects of the intervention are documented, an instrument to communicate them to respondents must be designed. Real world respondents may come to CV exercises with a great deal of information or no knowledge at all regarding the relevant attributes of the environment.[3] How much knowledge they have prior to the survey must be considered and perhaps assessed in advance through qualitative research. For respondents to be well informed, the knowledge they bring to the CV exercise may need to be augmented with information provided in the scenario.

All else equal, the communication burden placed on the CV scenario will likely be less when respondents have experience-based prior knowledge, than when their prior knowledge was based on media accounts and hearsay. Accordingly, studies that can build their scenarios on experiential knowledge will have the easiest time establishing their content validity. Those that must start from a very limited or non-existent knowledge base will have the most difficult cases to make.

In recognition of the importance of this aspect, the rating form allows up to 10 points to be assigned depending on how well the study documented and communicated

the potential effects of the intervention.

(4) Were respondents aware of their budget constraints and of the existence and status of environmental and other substitutes?

Because true values are defined in a framework involving budget-constrained utility maximization, many, including the NOAA Panel, argue that study subjects ought to be explicitly reminded of their budget constraints. Failure to do so would reduce the content validity of a study in the eyes of many potential reviewers.

Thus far, only the elements of the vector Q that would be affected by the intervention have been considered. Theory tells us that the value of environmental amenities affected by the intervention may depend on the status of other amenities that are substitutes for the potentially affected ones. Content validity may, therefore, be enhanced by assessing respondents' knowledge of the existence and status of substitutes during qualitative research and, if necessary, adding information about substitutes to the scenario. Furthermore, the range of substitutes may extend beyond environmental substitutes and include other public and private goods. Presumably complements should also be considered, but there is less emphasis on them in the thinking of many scholars, including members of the NOAA Panel.[4]

Figure 1 recommends up to 5 points be awarded, depending on the reviewer's judgement as to whether subjects were cognizant of their budget constraints and well informed about substitutes.

(5) Was the context for valuation fully specified and incentive compatible?

In addition to providing respondents with needed information about the effects of the intervention, a CV scenario will normally provide them with what we shall term the "context for valuation." Context refers to all dimensions of the proposed transaction dealing in one way or another with the how decisions about the intervention will be made and how money referred to in the CV question will be transferred. Whether the money will be paid to or received by respondents needs to have been clearly spelled out. Points might be lost, for example, if the nature of the value to be expressed was vague (e.g., asking "What is it worth to you?"). Whether the value is to be that of the individual or of the household needs to be clearly stated. Who else will be paying or receiving payment (the so-called "extent of the market," see Smith) may matter for environmental amenities with public goods characteristics. Certainly, theory dictates that the timing of payments has relevance to valuation. A valid CV study will strive to make the context of valuation as complete as possible.

Furthermore, theory raises some rather stern warnings about the incentive properties of CV scenarios. Incentive compatibility of payment mechanisms is an issue even for amenities, such as recreational opportunities, with private-good characteristics. It is well known, for example, that sealed-bid auctions create incentives to bid less than one's maximum willingness to pay, whereas a Vickery auction should lead to full value revelation, all else equal. This theoretical result may have practical relevance to studies using an open-ended CV format.

Where environmental amenities take on public-good characteristics, incentive issues are magnified because of the possibility of free riding and strategic responses. The theoretical strength of the referendum format in this context are widely accepted (e.g., Mitchell and Carson and Hoehn and Randall) and led the NOAA Panel to advocate heavy reliance on referenda in CV studies for purposes of damage assessment. In such circumstances, use of referendum formats, as opposed to voluntary donations, for example, would enhance content validity in the eyes of many reviewers. In our weighting scheme, if the context for valuation is complete and fully incentive compatible, it would be awarded 10 points. Studies with incomplete contexts would fare less well. Fewer points would also be assigned to studies with scenarios that are incentive incompatible in recognition of the potential confusion or strategic responses that such scenarios might induce.

(6) Did survey participants accept the scenario? Did they believe the scenario?

CV researchers and others (e.g., the NOAA Panel) have come to recognize that it is important that the scenario not only be communicated effectively, but that respondents accept it. A study subject accepts the scenario when he or she implicitly agrees to proceed with the valuation exercise based on the information and context provided. Scenario rejection can lead either to poor quality valuation data or item non-response for CV questions.

Content validity would be enhanced if respondents not only accept the scenario,

but believe it. Those writing on CV often emphasize that it involves "hypothetical" valuation, but some scenarios are more hypothetical than others. In many settings, asking study subjects to play "what if" games in order to value the intervention is unavoidable because a fully believable scenario is impossible to construct. However, in some circumstances, it may be possible to construct a scenario with a high degree of plausibility.

An example from the author's current research will illustrate. The work focuses on possible modifications in how Glen Canyon Dam on the Colorado River is operated. Changes may be needed to protect and enhance resources downstream in the Grand Canyon. Modifying dam operations would reduce its ability to generate electricity on-peak. A very likely result will be increases in how much some households in several western states will pay for electricity. One sampling frame for the CV study on this problem is the potentially affected electricity consumers. A referendum format is being used and the payment vehicle for this sampling frame will be electricity costs to these households. Focus groups showed that subjects found it very plausible that they would have to pay more for electricity if dam operations are modified. This enhances the credibility of their responses.

The rating form suggests that reviewers assign up to 10 points depending on their evaluation of whether respondents accepted the scenario and whether they found it believable. To earn all 10 points a study would have to demonstrate rather unambiguously that respondents both accepted and believed the scenario. We would

personally assign a fairly high score, perhaps 7 or 8, to a study that was very forthright about the hypothetical nature of the valuation exercise (thus foreclosing belief), but showed clear evidence that respondents nevertheless accepted the scenario. Whether respondents accepted and believed the scenario is admittedly difficult to determine, but some evidence can often be mustered. After careful consideration of the instrument, reviewers will no doubt form judgements about the plausibility of the scenario and the potential for scenario rejection. Furthermore, whether potential respondents accept and believe the scenario can be intentionally evaluated during focus groups and other procedures followed during qualitative phases of the research. Reports of such activities may help to reviewers evaluate these two dimensions. Furthermore, debriefing questions may be included in the survey to help determine rates of acceptance and belief.

(7) How adequate and complete were survey questions other than those designed to elicit values?

CV surveys typically include many questions other than those intended to elicit values. Several different objectives may be involved. For one, CV researchers often find it desirable to investigate respondents' motives for answering CV questions as they did. The exact form of such questions depends on both the form of the CV question and the researcher's judgement. For example, open-ended CV questions are often followed by questions designed to tell explore what respondents intended when they responded with a zero. A respondent may actually have had a zero value for the intervention, but a zero may also have been intended to communicate that the respondent did not know her

value, refused to place values on the intervention, rejected the scenario, or hoped that her response would reduce fees actually paid. The NOAA Panel, which, as noted already, recommended that a referendum format be used, also recommended that voting be followed by a question in an open-ended format asking respondents to explain why they voted as they did.

Additional questions may be included in the survey to provide evidence of its content validity. For example, appropriately worded questions could help evaluate whether respondents understood descriptive material in the scenario. Many past studies have included follow-up questions to attempt to identify strategic responses.

Other questions may also be included to assess the construct validity of the study. Construct validity tests normally involves hypotheses about relationships between answers to CV questions and other variables either in cross tabulations or in multiple-regression analyses (Bishop et al. 1994). Many types of questions can be included in the survey to support such analyses. For example, the NOAA Panel recommended cross tabulations of valuation responses with income, knowledge of the site, prior interest in the site for visitation or other reasons, environmental attitudes, attitudes toward big business, distance of residence from the site, understanding of the valuation task, and willingness and/or ability to perform the task.

Such survey questions need to be scrutinized as part of content validity assessment. Only if they are well designed will responses provide supporting data needed to meet the various objectives just noted. Because such questions are so important for construct

validity testing and other purposes, the rating form assigns 10 points to this dimension.

(8) Was the survey mode appropriate?

Mail surveys are attractive to CV researchers because they are the least expensive of the major modes. There also may be methodological reasons for choosing a mail approach. Mail is preferred by some researchers because mail instruments give them complete control over the information and context communicated to potential respondents. Other researchers shy away from mail surveys because of limited reading skills of potential respondents from the general population, even in the US and other countries where literacy rates are relatively high. Furthermore, even the more literate respondents may be reluctant to try to read and digest large amounts of written material about the intervention and its consequences.

Telephone interviews are more expensive than mail surveys and are limited in the amount of information and context that can be communicated during a brief phone call. Effective communication may require presenting respondents with visual aids such as charts, graphs, and photographs. This will not be feasible in a survey conducted entirely by phone. On the other hand, it is somewhat easier to get reasonably high response rates by phone than by mail and reading skills are not involved.

Personal interviews can make communication easier because of the personal contact between respondent and interviewer. More information can normally be provided than would be possible by mail or over the phone. Conducting surveys in person may

increase response rates. However, in-persons surveys with high response rates are very expensive. Likewise, the presence of an interviewer may influence responses.

From the perspective of content validity assessment, survey mode must be appropriate for the study goals and the complexity of the information and context that need to be communicated. If the goal is to value a recreational experience that is quite familiar to respondents, for example, then a mail survey may be quite adequate. If the goal is to estimate non-use values for a spill that had complex impacts on environments unfamiliar to respondents, then, as the NOAA Panel recommended, personal interviews would appear to have a large advantage. Using a mail or telephone survey in such a situation would be grounds for questioning the content validity of a study. This is not to say that a mail or telephone survey would necessarily be ruled out. However, in the eyes of many CV researchers, an extra burden of proof would rest on the study team to provide evidence that the mail or telephone procedures worked well.

Many CV researchers stress the importance of survey mode and we agree by assigning up to 10 points to this item.

(9) Were qualitative research procedures, pretests, and pilots sufficient to find and remedy identifiable flaws in the instrument and associated materials?

Once survey designers have roughed out an instrument and related documents based on their understanding of how respondents will react, qualitative research is often needed to refine the instrument. For example, focus group participants may be asked to complete a draft mail survey and then discuss it with the group leader. Or, an

instrument designed for personal interviews can be tested in observed interviews. During such interviews, and afterwards in debriefing sessions with the subjects, researchers can try to identify ways that the instrument is being misinterpreted or if information provided is incomplete or otherwise inadequate. Possible improvements can be tested as well. Qualitative testing should not only involve verbal materials but also any photographs or other visual aids.

Formal pretesting and piloting[5] of a nearly finished instrument may also improve it. Statistical analyses of responses provides a preview of what to expect in the final results and can help diagnose problems. Interviewers often help to identify places where in-person and telephone questionnaires can be improved. Interviewers can also be instructed to record verbatim any remarks by respondents about the survey questions and information presented. Though less effective, subjects in mail pretests and pilots can be asked to write comments in the margins. A subsample can be contacted by telephone to probe for flaws in a draft mail instrument. Through such procedures, the study design can be tested under field conditions, enhancing content validity in the process.

Though we share the now commonly accepted view that qualitative research can be invaluable in the design of CV surveys, its limitations in supporting validity must also be recognized. The typical study report will include only a terse statement such as, "Four focus groups were conducted." Little or nothing is said about the extent to which the focus groups succeeded in working the "bugs" out of the instrument and associated documents. Standard procedures for applying qualitative research tools and reporting the

results do not exist, or at least have not found their way into everyday practice in economics. This may be a fruitful area for research. In the meantime, reviewers of CV studies may have to take the "quality" of qualitative work more or less at face value. An exception may be litigation, where details about procedures and results can be ferreted out from audio and video records, from written reports entered into evidence, and from depositions and cross examination.

CV can be applied in such diverse settings that generalizations are not possible regarding how much qualitative research, pretesting, and piloting are needed in any particular case. At one extreme are studies of relatively straightforward interventions, where there is a long history of past research upon which to draw. In such cases, instruments may require little preliminary testing. At the other extreme are non-use studies involving environmental resources unfamiliar to large numbers of potential respondents. Hence, judgements about the appropriate amount of preliminary work must take specific circumstances into account. Up to 5 points are to be assigned to this aspect under our version of the rating form.

(10) Given study objectives, how adequate were procedures employed to choose study subjects, assign them to treatments (if applicable), and encourage high response rates?

Adequate population definition, sampling, and survey procedures depend on study objectives. To allow for this fact, we will distinguish between two different kinds of studies. Some studies involve exclusively methodological goals. One might, for example, design a study to compare the results of open-ended CV questions with those

from a bidding game for the same amenity. Other studies have as a major goal the estimation of values for a population of individuals, either in the context of policy analysis or litigation. For convenience, we will term the former "methodological studies" and the latter "applied studies." Applied studies may also have methodological goals. Their distinguishing feature is that they have the ultimate goal generalizing results from sample to population.

For methodological studies, procedures for choosing subjects and allocating them among treatments are mostly a matter of common sense. Where new CV procedures or hypotheses about CV data are to be tested, one would hope to eventually conclude something about how CV would perform in applied studies under normal circumstances. Hence, one might not want to choose kindergartners as subjects. Content validity might suffer a bit if only undergraduates were used as subjects since their responses might be very different from general population samples used in many CV studies. However, at the other extreme, fastidious sampling from the general population or some other group would normally not be required for methodological studies. If the goals of the research are purely methodological, the self-selection bias inherent, for example, in recruiting from the general population subjects who are willing to come to a laboratory and participate in an experiment would probably not be a large red flag in most researchers' and reviewers' judgement. In studies involving multiple treatments, assignments to cells should, of course, be random. In field (as opposed to laboratory) studies, follow-up procedures to increase response rates could normally be less rigorous than in an applied

study. In sum, the validity of implementation steps for methodological studies focus mainly on the reasonableness of the procedures in light of the study goals.

Applied studies, on the other hand, must satisfy more rigorous standards as far as sampling and response rates are concerned. Either random or stratified random samples are required which will support extrapolation of value estimates from sample to population. Furthermore, potential non-response bias must be addressed. The best way to head off non-response bias is by gaining a high response rate in the first place. Survey researchers have well developed procedures for doing so. Various methods to gain a rough idea of the potential seriousness of non-response bias are available. An example would be to compare reported socioeconomic characteristics of respondents with published statistics for their Census tracts. In some cases, population statistics are available in sufficient detail to allow weighting of the sample to represent the population. Careful attention to this issue enhances content validity.

Up to 10 points can be allocated to a study depending on how well it dealt with sampling, non-response, and related details within the context of its overall objectives.

(11) Was the econometric analysis adequate?

Once the responses are in, high content validity requires that the data be competently coded and entered into computer files for analysis. Success here again is simply a matter of using common sense. For example, verification of data is often facilitated by entering it twice and reconciling the data files.

The analysis itself should employ econometric procedures that are appropriate to the data and the inferences that are to be drawn. Economists are normally well trained in this area. Assessing this aspect of content validity is mostly a matter of verifying that analysts have employed their tools properly. We assign 10 possible points to this aspect.

(12) How adequate are the written materials from the study?

The final step in study execution involves reporting study design and execution procedures and study results. Needs here will vary depending on study goals and the expected audience for the report. A journal article might stress technical and methodological details, while a report for policy makers might stress final results and policy implications. Study reports should reflect such objectives.

Content validity assessment itself requires rather complete reporting. Because the burden of proof for content validity rests with the researchers, studies that do not provide thorough and complete reports can not be presumed to have high content validity. This no doubt was part of the motivation for the NOAA Panel's rather severe requirements for reports:

> Every report of a CV study should make clear the definition of the population sampled, the sampling frame used, the sample size, the overall sample non-response rate and its components (e.g., refusals), and item non-response on all important questions. The report should also reproduce the exact wording and sequence of the questionnaire and of other communications to respondents (e.g., advance letters). All data from the study should be archived and made available to interested parties . . .

From the somewhat broader perspective taken in this paper, the ideal study report would

also include a clear statement of the study goals, a definition of the true value to be estimated, a description of the intervention and its effects on environmental amenities, and a fairly detailed summary of the procedures followed throughout the study. The rating form asks reviewers to assign up to 5 points for this aspects.

III. OVERALL EVALUATION QUESTIONS

(13) Total Points

Once the detailed study procedures have been scored, the rating form suggests that the reviewer add up the points. Some reviewers may wish to skip this step, arguing that it implies a degree of quantitative precision far beyond what can be hoped for under the current state of the art in CV. We can certainly appreciate the reasons for such a reservation. We would nevertheless encourage reviewers to struggle with the numbers, including their aggregate value. We believe that doing so will promote balance in appraisals of content validity. In considering such a complex set of issues, one may tend to focus too much attention on some aspect that seems particularly well done or innovative, or on some flaw that is particularly glaring. Without the discipline imposed by assigning and summing the numbers, too little weight may implicitly be assigned to other study procedures that were done well or poorly. Struggling with the numbers and aggregating them will help avoid such imbalances. Furthermore, it may encourage deeper consideration of the criteria themselves. Particularly after several applications of the rating form, one may feel that the score for a given study seems too high or too low.

If so, this may indicate that the weights on the individual items are not in keeping with that reviewer's more fundamental judgements about the relative importance of the various issues raised in the individual detailed questions. The weights may need to be adjusted. In the process of considering this issue, reviewers can force themselves to more carefully consider the criteria they apply and the relative importance they place on different criteria.

(14) Are there other concerns relating to the design and execution of the study that have not already been addressed?

At this point, before the final step in the rating process, we confront two problems. First, CV study procedures still involve many dimensions about which widely-respected researchers disagree. There may well be dimensions that some feel are essential that are not even mentioned here. Second, Question 14 will come into play when special circumstances not ordinarily faced in CV studies are present. For example, timing of survey administration may be an issue in some circumstances but not in others. Suppose injuries due to a large oil spill are to be valued. Doing a CV study too soon afterward might be challenged on the grounds that respondents were still in a state of shock and outrage, and answered the survey in ways that reflected emotions of the moment. Resulting value estimates would be of questionable validity because they might not be robust over time.

Question 14 provides the opportunity for reviewers to write in concerns and issues not raised elsewhere in the rating form, including those that were more or less unique

to the particular study being reviewed.

The final step in the content validity assessment is to sum up the reviewers overall evaluation of the study by responding to Question 15.

(15) Considering the issues raised in Questions 1 through 12, your total score as calculated for Question 13, and any additional issues raised under Question 14, how would you rate this study overall?

_____ Excellent
_____ Good
_____ Fair
_____ Poor
_____ Unacceptable (Study Fatally Flawed)

The response to this question should help interpret the numerical scores. Suppose, for example, that a study received an aggregate score of 50 points. Such a score would almost certainly be inconsistent with a rating of "excellent" or even "good," but would not convey whether the study was judged "fair," or "poor" or even "unacceptable." The qualitative rating in the final question should help to clarify how serious the potential flaws in the study were judged to be.

A rating of "unacceptable" would signify that a study had fatal flaws. This response would be appropriate if the study failed to meet the reviewer's minimum standards. Suppose, for example, that a study employed telephone interviews in a way the reviewer judged to be not at all adequate to provide sound CV data. Such a study would fail to meet this reviewer's minimum requirements under Question 8. The reviewer would declare the study unacceptable under Question 15 regardless of the total

points it earned when the detailed question scores were added. A study that failed to communicate well, neglected to provide a minimally adequate context, or failed miserably elsewhere should simply be identified as unacceptable.

The link between study goals and the criteria for fatal flaws is important to remember. A study designed to be a first preliminary investigation of benefits or natural resource damages, for example, should not be held to the same standards as one that is designed to serve as a basis for an important policy analysis or a final damage estimate. A low-budget study designed to serve primarily as a student project might leave many loose ends that would be unacceptable in a study destined to be used to set damage in an important court case.

IV. SUMMARY AND SOME FINAL THOUGHTS

In this paper, we have attempted to clarify and systematize an approach to content validity assessment for CV studies. A content valid CV study is rooted throughout in a clear theoretical definition of the true value of the intervention. At the heart of such a study will be its scenario. Based on well-documented evidence of the respondent-relevant effects of the intervention, a sound scenario effectively communicates the potential effects of the intervention to respondents. It includes whatever information they need regarding substitutes for the environmental resources in question and may need to remind them of their budget constraints. It also includes a fully specified and incentive compatible context for valuation. It does all this in ways that potential respondents will accept and,

if possible, believe.

Looking beyond the scenario, a content valid survey instrument will include well-designed questions to support construct validity testing and achieve other goals. The mode chosen for administering the survey will be appropriate to the complexity of the scenario and the ultimate goals of the study. Prior to administration, the instrument will have been subjected to sufficient qualitative investigation, pretesting, and, if needed, piloting to work out as many bugs as possible. Econometric analysis of the results will have been adequately performed and final results effectively reported.

When studies fall short of these ideals, as nearly all will, they may still have substantial merits. Content validity is normally a matter of degree. However, some studies will fall below minimal standards and be judged content invalid. Other studies, though they may not be rejected outright, may still be viewed with substantial reservations because of possible flaws in design and execution.

To admit evidence from surveys into applied welfare studies, where revealed-preference data have historically dominated, is a big step for many economists. Whether contingent values ought to be considered "admissible evidence" should be approached in a cautious, but open-minded, way based on carefully thought out "rules of evidence." Thus do the social sciences progress. Drawing on its sister disciplines, economists can evaluate this new direction based on content, construct, and criterion validity. Content validity deserves more attention if real progress is to be made.

Figure 1

CONTENT VALIDITY RATING FORM FOR CONTINGENT VALUATION STUDIES

(1) Was the theoretical true value clearly and correctly defined? (5 points) . . . . . _____

(2) Were the environmental attributes relevant to potential subjects fully identified (10 points) . . . . . _____

(3) Were the potential effects of the intervention on environmental attributes and other economic parameters adequately documented and communicated? (10 points). . . . . _____

(4) Were respondents aware of their budget constraints and of the existence and status of environmental and other substitutes? (5 points). . . . . _____

(5) Was the context for valuation fully specified and incentive compatible? (10 points). . . . . _____

(6) Did survey participants accept the scenario? Did they believe the scenario? (10 points). . . . . _____

(7) How adequate and complete were survey questions other than those designed to elicit values? (10 points) . . . . . _____

(8) Was the survey mode appropriate? (10 points). . . . . _____

(9) Were qualitative research procedures, pretests, and pilots sufficient to find and remedy identifiable flaws in the instrument and associated materials? (5 points). . . . . _____

(10) Given study objectives, how adequate were procedures employed to choose study subjects, assign them to treatments (if applicable), and encourage high response rates? (10 points) . . . . . _____

(11) Was the econometric analysis adequate? (10 points). . . _____

(12) How adequate are the written materials from the study? (5 points). . . . . _____

(13) TOTAL POINTS:. . . . . _____

Figure 1 (continued)

(14) Are there other concerns relating to the design and execution of the study that have not already been addressed? _____

_____

_____

_____

_____

(15) Considering the issues raised in Questions 1 through 12, your total score as calculated for Question 13, and any additional issues raised under Question 14, how would you rate this study overall?

_____ Excellent
_____ Good
_____ Fair
_____ Poor
_____ Unacceptable (Study Fatally Flawed)

REFERENCES

American Psychological Association. 1974. *Standards for Educational and Psychological Tests*. Washington, D.C.

Bishop, Richard C., Patricia A. Champ, Thomas C. Brown and Daniel W. McCollum. "Non-Use Values: Theory and Empirical Applications." Paper presented at the Werner Reimers Meeting on Determining the Value of Non-Market Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods, Bad Hamburg, Germany, July 24-29, 1994, proceedings forthcoming.

Bishop, Richard C., and Thomas A. Heberlein. 1979. "Measuring Values of Extramarket Goods: Are Indirect Measures Biased?" *American Journal of Agricultural Economics* 61(5):926-930.

Bishop, Richard C., and Michael P. Welsh. 1992. "Existence Values in Benefit-Cost Analysis and Damage Assessment." *Land Economics* 68(4):405-417.

Bishop, Richard C., Michael P. Welsh, and Thomas A. Heberlein. 1993. "Some Experimental Evidence on the Validity of Contingent Valuation." Unpublished paper, Department of Agricultural Economics, University of Wisconsin--Madison.

Bishop, Richard C., Patricia A. Champ, and Daniel J. Mullarkey. 1995. "Contingent Valuation." In *The Handbook of Environmental Economics*, Daniel W. Bromley, Ed. Cambridge: Basil Blackwell Ltd. pp. 629-654.

Bohrnstedt, George W. 1983. "Measurement." In *Handbook of Survey Research*. Peter H. Rossi, James D. Wright, and Andy B. Anderson, Eds. New York: Academic Press. pp.69-121.

Carson, Richard T., Robert C. Mitchell, W. Michael Hanemann, Raymond J. Kopp, Stanley Presser, and Paul A. Ruud. 1992. "A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill." Report Submitted to the Attorney General of the State of Alaska. Reprinted by Natural Resource Damage Assessment, Inc.

Cummings, Ronald G., David S. Brookshire, and Willam D. Schulze, eds. 1986. *Valuing Environmental Goods: A State of the Art Assessment of the Contingent Valuation Method*. Totawa, N.J., Rowan and Allanheld.

Hanemann, W. Michael. "Valuing the Environment Through Contingent Valuation,"

Journal of Economic Perspectives 8(4):19-44 (Fall 1994)

Hausman, Jerry A., Ed. *Contingent Valuation: An Assessment*. Amsterdam: North-Holland Publishing, 1993.

Hoehn, John P., and Alan Randall. 1987. "A Satisfactory Benefit Cost Indicator from Contingent Valuation." *Journal of Environmental Economics and Management* 14(3):226-47.

Loomis, John, Armando Gonzalez-Caban and Robin Gregory. 1994. "Substitutes and Budget Constraints in Contingent Valuation." *Land Economics* 70(4):499-506.

Mitchell, Robert C., and Richard T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Washington, D.C.: Resources for the Future.

Schkade, David A., and John W. Payne. 1994. "How People Respond to Contingent Valuation Questions: A Verbal Protocol Analysis of Willingness to Pay for an Environmental Regulation." *Journal of Environmental Economics and Management* 26:88-109.

Smith, V. Kerry. 1993. "Nonmarket Valuation of Environmental Resources: An Interpretive Appraisal." *Land Economics* 69(1):1-26.

Sundberg, N.D.· 1977. *Assessment of Persons*. Englewood Cliffs, New Jersey: Prentice Hall, Inc.

U.S. Department of Commerce. National Oceanic and Atmospheric Administration. 1993. *Natural Resource Damage Assessments under the Oil Pollution Act of 1990*. Federal Register 58(10):4601-4614.

Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the Social Sciences: The Link Between Theory and Data*. Cambridge: Cambridge University Press.

Footnotes

1. Our definition of content validity assessment is significantly broader than that of Mitchell and Carson (pp. 190-192). They focus exclusively on examination of the survey instrument, while we would include all aspects of study design and execution.

2. That is, those who, as journal reviewers, consultants to decision makers, expert witnesses, or in other such roles, are called upon to evaluate the merits of CV studies.

3. We will not consider here the ongoing debate among environmental economists about whether the status of an attribute can be "relevant" to consumers who are not aware of it. For one view that has found its way into print, see Bishop and Welsh. Basically, that paper argues that, as a practical matter, real world consumers can not be expected to have full knowledge about all the things affecting their welfare. Obscure and even unknown environmental resources could have value to them.

4. The extent to which it is necessary to explicitly deal with budget issues and substitutes in CV scenarios remains a subject for further research. At least one published study (Loomis, Gonzalez-Caban, and Gregory) has found statistically indistinguishable results whether budget constraints and substitutes were mentioned or not.

5. Pretests are distinguished from pilots by their small and more convenient samples. The goal of pretests is to identify major problems with of survey design that will be apparent even for small samples. Pilot studies are conducted to further refine question and information wording, test proposed procedures for the final survey under field conditions, and investigate the likely statistical properties of final results.

ADDENDUM ONE

Addendum 1

Responses to Criticisms of My Peer Review
by ARCO's Experts

Two economic experts for ARCO, Dr. William H. Desvousges and Professor Jerry A. Hausman, criticized my peer review of the State of Montana's Contingent Valuation study. In this Addendum, I would like to respond briefly to their criticisms.

Dr. Desvousges (p.62): "Professor Bishop's involvement in the State's study compromises his objectivity as a peer reviewer."

Dr. Desvousges suggests that because I was an outside consultant when the Montana CV study was being conducted, I was not objective in reviewing the final report. Peer reviews vary in terms of procedures. Dr. Desvousges mentions formal reviews of research papers for possible publication in scientific journals as one model, but other models are applied. The journal-review model can have the unfortunate result that flaws are identified in the research that could have been corrected early on had the researchers known about them. For this reason, applied studies done for policy analysis and litigation as well as efforts to write scientific books and textbooks sometimes use peer reviewers as projects evolve. Through mid-course corrections, researchers and managers hope to have a stronger final product and to avoid finding out about flaws only when it is too late.

My participation in the design of the Clark Fork study had this goal. As Dr. Desvousges points out, I was an outside consultant. To the extent that I participated, I was involved in the research process through telephone conversations with the investigators. I had no decision making authority over any aspect of the study, nor did I analyze any data, attend any focus groups, or otherwise participate directly in the research process in any way

that compromised my objectivity. I simply offered my thoughts as a reviewer during the process, advice that the investigators could accept or disregard as they chose. The sorts of more intimate participation that could have led to loss of objectivity did not occur. To the extent that I was somewhat familiar with the work prior to undertaking my review of the final report, my formal review was more comprehensive than it would have been otherwise.

Dr. Desvousges (p.63): "Professor Bishop's review ignores the one test most directly related to the validity of a contingent valuation study."

Dr. Desvousges bases this criticism on a now outmoded view which I, in earlier days, helped to promote. I no longer agree that a few simulated market studies conducted under very specific conditions can form a sound basis for deciding whether or not the contingent valuation method is valid in general. I and others who promoted this view were rather naive. Each individual study has so many unique aspects, some of which are obvious and some of which are likely to be latent, that generalizations based on a few studies are not likely to be valid. That is a fundamental point of my section on the "Overall Validity of the Contingent Valuation Method" and the paper included in Appendix A of my peer review.

Dr. Desvousges (p.64): "Professor Bishop's procedure for assessing validity is subjective, untested and experimental."

My approach to validity does depend on the professional judgement of the reviewer, as I have stressed in my review and its appendices and in my deposition. I do not believe that this makes it any more "subjective" than most of the other procedures we apply to

evaluate the validity of studies in the social sciences.

My approach to validity is not revolutionary. The "Three Cs" of content, construct, and criterion validity are standard in the social sciences. While they are somewhat less familiar to economists, they are discussed in detail in the book on contingent valuation by Mitchell and Carson (1989). I am merely trying to build on their work on the topic. Nor does one find anything particularly original in the individual criteria I apply. To the extent that my work on validity makes an original contribution, it does so merely by making generally accepted principles for assessing the validity of contingent valuation studies a little clearer and more systematic.

Dr. Desvousges (p.64) argues that my criteria "differ markedly from criteria proposed by other economists and contingent valuation practitioners. (See Cummings, Brookshire, and Schulze, 1986, p. 104; Mitchell and Carson, 1989, p. 192; 58 Fed. Reg. 4608-4609.)" Consider the three cited works for a moment. I have always found the four "Reference Operating Conditions (ROCs)" of Cummings et al. very incomplete. The four ROCs taken together barely scratch the surface so far as the range of issues that arise in trying to evaluate the performance of contingent valuation studies. Nothing is said about theoretical issues, the importance of identifying respondent relevant attributes, the role of budget constraints and substitutes, or most of the other issues raised on the Content Validity Rating Form. Nor is there any mention of the desirability or interpretation of construct validity test or criterion validity studies.

As for Mitchell and Carson (1989), Dr. Desvousges refers to a two paragraph discussion of content validity that appears on their page 192. Contrary to what Dr.

Desvousges seems to be trying to imply, the overlap between what they advocate and my criteria is striking. They mention the importance of accurate descriptions of the environmental amenity to be valued, the need to make the scenario consistent with relevant theory, the need for care in instrument design, and the importance of adequate reporting in ways that are very similar to mine. They then add the following paragraph (Mitchell and Carson 1989, p. 192),

> The kinds of questions appropriate to an assessment of a scenario's content validity include: Does the description of the good and how it is to be paid for appear to be unambiguous? Is it likely to be meaningful to the respondents? Is there anything in the scenario that might suggest to some respondents that the good would not be paid for? Are the property rights and market for the good defined in such a way that the respondents will accept the WTP format as plausible? Does the scenario appear to force reluctant respondents to come up with WTP amounts? Although the answers to these questions are necessarily subjective and open to debate, this type of assessment should be undertaken whenever CV-based estimates are used for policy purposes.

There is a close parallel between almost every one of these questions and my procedures.[1] It is true that Mitchell and Carson chooses to discuss issues such as survey mode and econometric analysis elsewhere, rather than at this point, but most if not all of the points that are dealt with in the Content Validity Rating Form are covered in their book. How Dr. Desvousges can conclude that the approach advocated by Mitchell and Carson "differs markedly" from mine is beyond me.

The citation to the Federal Register refers to the report of the NOAA Panel on Contingent Valuation. The overlap between their guidelines and the issues I identify is

---

[1] The one exception concerns possible coercive questioning. I take it for granted that questions forcing "reluctant respondents to come up with WTP amounts" would be a red flag, whereas they choose to mention it explicitly.

large. Consider some examples. The NOAA Panel advocated personal interviews. I disagree up to a point, but certainly considered the same issues under Question 8 of the Content Validity Assessment Form. The NOAA Panel advocated careful sampling and so do I (Question 10). The NOAA Panel stressed the need for high response rates, and so do I (also under Question 10). Similar concerns motivated the NOAA Panel's emphasis on reporting and my own in Question 12. The NOAA Panel advocated careful attention to providing respondents with accurate information, a need that I stress under Question 2. The NOAA Panel advocated that other questions be included in the survey to help interpret results, as I do under Question 7. The panel placed great importance of exploring the relationships between responses to valuation questions and responses to these other questions, a need I cover under construct validity testing. The NOAA Panel adocated reminding respondents about budget constraints and substitutes, as I do in Question 4. Though I do disagree with the NOAA Panel on some of the fine points, I again fail to see what Dr. Desvousges has in mind when he says that the criteria I advocate "differ markedly" from those of the NOAA Panel.

<u>Dr. Desvousges (p. 66): "Professor Bishop's endorsement of the state's contingent valuation study is lukewarm."</u>

If by "lukewarm" Dr. Desvousges means that I found a number of questions to raise about the Clark Fork study then he is right. The Clark Study is not perfect. It is incumbent on peer reviewers in my position to call attention to whatever flaws they find in a study. Once those flaws are on the table it is then necessary to ask how serious they are and

whether or not the study is of sufficient quality to serve the purpose it was designed to serve. It would be most inappropriate to become so bogged down in the individual potential shortcomings I identified that my overall conclusion from the review, expressed in its very last sentence, were to get lost: "In my opinion, then, results of the Clark Fork study have sufficient validity to be used as measures of the true values of Montana residents for partial and complete cleanup of the Clark Fork sites." Continuing Dr. Desvousges' metaphor, while I found enough issues during my review to make me less than "red hot" for the Clark Fork study, my review turned up so many positive features that I feel substantial "warmth" for it.

Dr. Desvousges (p. 67): "Professor Bishop agrees that the State's study fails the most rigorous scope test."

I continue to believe that the Clark Fork study's failure to pass a between-sample scope test when using the full samples for Versions 1 and 2 is a mark against it. The question is how much to make of this failure. It is important not to forget that scope tests were passed at conventional levels of significance for one major subsample, as discussed in Hagler Bailly's rebuttal report, Chapter 4. Small sample sizes are most likely to blame for lack of significance for another subsample, the "site specific" group. The nearly significant difference in values for a subsample in the original report combined with these new results do much to allay concerns I had about the issue.

It is also my belief that within-sample scope test have more potential relevance than Dr. Desvousges and Professor Hausman are willing to admit. Scope tests per se only took on the importance they now have in everyone's minds after publication of the NOAA Panel

Report in 1994. The NOAA Panel focused on between-sample designs and I had not paid much attention to the possibility of within-sample scope tests prior to becoming involved in the Clark Fork study review. That review prompted me to consider the possible merits of within-sample tests. Though somewhat weaker than between-sample scope tests, within-sample scope tests seem to me now to have considerable potential. If a study fails a within-sample scope test, this would surely raise doubts about its validity. On the other hand, if a study passes a within-sample scope test, as the Clark Fork study did, then this is a positive sign although some ambiguity of interpretation remains. One interpretation which ARCO's experts have seized upon is that study subjects might have taken cues from seeing both of the scenarios together about how the survey designers wanted them to respond. However, so far there is no evidence one way or the other about whether such cues have impacts or how strong they are. Alternatively, perhaps some respondents need to see the alternatives back-to-back to sort out their preferences. Prior to further research to resolve what actually happens in within-sample scope tests, it would be premature to dismiss them out of hand based on unproven, negative interpretations about how respondents might react to multiple valuation questions. That the Clark Fork study passed within-sample scope tests and was able to pass between-sample scope tests across some subsamples ought to be interpreted as positive evidence regarding the scope issue, although it remains somewhat less convincing than passage of a full between-sample scope test would have been.

Professor Hausman (p. 53): "If the CV method itself is unreliable, a 'well-done' CV study (one with high content validity) would not be markedly more reliable than a poorly done CV

study. Thus, the content validity criteria are weak standards by which to judge a CV study."

Professor Hausman and I begin from different premises here. He is ready to reach rather sweeping negative conclusion about the contingent valuation method. I have concluded that contingent valuation studies can be sufficiently reliable to justify their use in policy analysis and litigation, although there are many issues yet to be researched.

How is it that Professor Hausman and I come to such different conclusions? My conclusion is based on a much larger and broader set of studies than Professor Hausman seems to be willing to consider. In fact, Professor Hausman is rather selective in choosing studies on which he proposes to base his conclusions about contingent valuation. I refer not only to his expert report for ARCO, but also his various contributions to his own edited volume (Hausman 1993) and his co-authored article (Diamond and Hausman 1994) in the Journal of Economic Perspectives. Much of the evidence that Professor Hausman uses comes from studies that were performed in preparation for litigation over the Exxon Valdez oil spill, studies that were designed from the beginning to discredit contingent valuation. These studies typically have rather low content validity, as my peer review shows in the case of the study of logging in wilderness areas, one of the studies on which Professor Hausman depends heavily. To the extent that he reaches out to studies other than those done for Exxon, Professor Hausman typically selects studies that show contingent valuation in the worst possible light. I have in mind studies cited in the Journal of Economic Perspectives article by Kahneman and Knetsch, Samples and Hollyer, Duffield and Patterson, Seip and Strand, and Neil et al. Like the Exxon studies, most of these studies would rate relatively low in content validity. Studies that might point to more favorable conclusions about

contingent valuation tend to be ignored when Professor Hausman evaluates contingent valuation.

If, despite his unwillingness to consider much of the literature, one accepts Professor Hausman's negative verdict about contingent valuation, using measurement techniques that display high content validity will not help. If contingent valuation never works, details of the procedures followed won't change that fact. On the other hand, under my belief that contingent valuation can provide useful information about preferences, it only stands to reason that better done studies ought to get more reliable results. This is the basic point of the NOAA Panel, as well, and its reason for dealing with study procedures in so much detail.

Professor Hausman (p.53): ". . . rudimentary construct validity tests, like content validity criteria, are very weak tests because they cannot distinguish between the theory that WTP answers reflect economic preferences and alternative theories that WTP answers are simply reflections of respondents' warm glow feelings or attitudes."

One problem with this criticism of construct validity testing is that the "alternative theories" that Professor Hausman wants to use to discriminate between "weak" and "strong" validity tests do not exist. There is no well-developed economic theory exploring how people's responses to contingent valuation questions are related to feelings of warm glow idea. Warm glow exists in the economics literature only as a theory of charitable giving, and not as a theory of how people respond to valuation questions in surveys. In the literature on contingent valuation, warm glow is nothing more than an ex post rationalization of the embedding phenomenon, a convenient story with a modicum of plausibility on a superficial level. No theoretical framework has yet been developed to undergird a theoretical distinction between valid values based on economic preferences and invalid values that are somehow based on the same sort of warm glow that is thought to motivate charitable giving. Perhaps that is why the NOAA Panel was not very concerned about warm glow arguments in its evaluation of contingent valuation. Professor Hausman also suggests a "theory" that contingent valuation responses are mere expressions of attitudes and thus should not be considered reliable measures of economic preferences. To my knowledge, attitudes have never been defined in mainstream economics nor have they been theoretically distinguished from economic preferences. Because warm glow and attitudes lack theoretical content, Professor Hausman's objection to construct validity tests lacks relevance.

From my perspective, we economists must rely primarily on our accepted body of theory at it currently exists. To the extent that competing theories, such as a theory of survey responses based on warm glow, are raised they must be carefully and rigorously worked out in a form that is empirically testable. Until such time as new theories displace old ones, the existing ones must continue to serve. Among other things, existing economic theory provides the basis for interpreting the results from value estimation studies. Value estimates from revealed preference and contingent valuation studies are interpreted as "indifference-producing amounts of money," as that term was developed in the body of my peer review and in its appendices. It is the chasm between observable data from markets and surveys and the unobservable mental state of indifference that theory helps us to bridge. To the extent that the relationships postulated in theory can be found in the market or survey data used to estimate economic values, confidence increases that those data were in fact generated by the processes and relationships that extant theory was designed to capture. This is the goal of construct validity testing. The larger the range of expected possible relationships that can be found in the data the more confident the researcher can be about interpreting results as economic values. When asking whether expected relationships are to be found in the data, one should not limit the range of tests to only a few such as a scope test or an additivity test. Hence, I believe that construct validity tests, including what I have termed rudimentary tests, are central to evaluating the validity of economic values estimated based on revealed preference and stated preference data, including contingent valuation studies.

References:

Cummings, R.G., D.S. Brookshire, and W.D. Schulze. 1986. *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method.* Totowa, N.J.: Rowman and Allenheld Publishers.

Diamond, Peter A., and Jerry A. Hausman. 1994. "Contingent Valuation: Is Some Number Better Than No Number?" *Journal of Economic Perspectives* 8(4):45-64.

Hausman, J., ed., 1993. *Contingent Valuation: A Critical Assessment.* Amsterdam: Elsevier.

Mitchell, R.C., and R.T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method.* Washington, D.C.: Resources for the Future.